МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Гидрофизики и гидропрогнозов

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

(магистерская диссертация)

На тему **Применение методов машинного обучения для прогнозирования гидрологических характеристик весеннего половодья**

Исполнитель Воронова Алёна Александровна
(фамилия, имя, отчество)

Руководитель кандидат географических наук, доцент
(ученая степень, ученое звание)

Шаночкин Сергей Владимирович
(фамилия, имя, отчество)

«К защите допускаю»
Заведующий кафедрой

(подпись)

К.т.н., доцент
(ученая степень, ученое звание)

Хаустов Виталий Александрович
(фамилия, имя, отчество)

«30» мая 2017 г.

Санкт–Петербург
2017

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра гидрофизики и гидропрогнозов

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

(магистерская диссертация)

На тему **Применение машинного обучения для прогнозирования гидрологических характеристик**

Исполнитель Воронова Алена Александровна
(фамилия, имя, отчество)

Руководитель кандидат географических наук, доцент
(ученая степень, ученое звание)
Шаночкин Свергей Владимирович
(фамилия, имя, отчество)

«К защите допускаю»
Заведующий кафедрой

(подпись)
кандидат технических наук, доцент
(ученая степень, ученое звание)
Хаустов Виталий Александрович
(фамилия, имя, отчество)

«___»_____2017 г.

Санкт–Петербург
2017

Содержание

Введение 4

1 Характеристика бассейна Северной Двины 6

1.1 Рельеф 9

1.2 Климат 10

1.3 Хозяйственное использование поверхностных вод 11

1.4 Описание поста 13

2 Методология нахождения поставленной цели 17

3 Семейства алгоритмов машинного обучения 21

4 Обработка исходных данных 22

4.1 Отбор признаков 23

4.1.1 Масштабирование данных 24

4.1.2 Кластерный анализ 25

5 Решающие деревья 32

5.1 Построение деревьев 33

5.2 Градиентный бустинг 37

5.2.1 Расчет характеристик весеннего половодья градиентным бустингом 40

5.3 Достоинства и недостатки решающих деревьев 53

6 Нейронные сети 55

6.1 Общая модель нейрона 56

6.2 Типы функций активации 57

6.3 Стохастическая модель нейрона 59

6.4 Нейронные сети прямого распространения 60

6.4.1 Расчет характеристик весеннего половодья при помощи многослойного персептрона 65

7 Основные результаты 74

Список литературы ..... **Ошибка! Закладка не определена.**

Введение

Условия, предшествующие весеннему половодью из года в год, могут сильно отличаться. Накапливается разное количество снега, тает он с различной интенсивностью, температурные условия могут сильно различаться, толщина льда так же меняется. На момент составления прогноза эти величины, как правило, известны, возникает вопрос-можно ли на основании столь разнородных данных дать качественный прогноз характеристик весеннего половодья? Такие прогнозы очень полезны для водопользователей. Известно, что многоводные дружные половодья приводят к наводнениям в населенных пунктах, прорыву плотин, спуску прудов, подтоплению сельскохозяйственных угодий и к другим негативным последствиям. Даже небольшие реки в период половодья могут наносить большой экономический и экологический ущерб.

В качестве объектов исследования выступают величины дат и расходов воды начала, пика и конца весеннего половодья Северной Двины. Предмет исследования – это методы машинного обучения, которые позволяют прогнозировать данные характеристики.

В настоящее время традиционные методики прогноза характеристик весеннего выдают результаты недостаточной точности, при этом, в большинстве случаев, сильно зависят от метеорологических прогнозов, которые имеют большую погрешность в долгосрочной перспективе. Цель работы – опробовать на практике некоторые методы из стремительно развивающейся области знаний, машинного обучения и оценить возможности применения этих методов в гидрологии.

Сама идея обработки гидрологических данных методами машинного обучения появилась довольно давно. Первые упоминания об этом можно найти в статье В.А. Румянцева [1] и позже в статье С.В. Шаночкина [2] . О математической реализации некоторых методов машинного обучения, уже был известно примерно в пятидесятые годы, вот только для реализации в то

время не хватало вычислительных мощностей, сейчас, когда вычислительные процессоры стали намного мощней, полученные знания стали воплощать в жизнь, что привело к новым открытиям в различных областях науки и техники.

Решаемые задачи:

1) поиск подходящих методов моделирования

2) подготовка исходных данных

3) построение предсказательных моделей

4) использование модели на практике.

1 Характеристика бассейна Северной Двины

Территория по которой течет Се́верная Двина́ , расположена на севере Европейской части России. Северная Двина образуется на месте слияния рек Сухона и Юг, неподалеку от Великого Устюга, на территории Вологодской области и течет в направлении Архангельской области. В районе города Котлас в реку впадает Вычегда, объем воды которой больше чем у Северной Двины, на участке от истока до впадения Вычегды Северную Двину ещё называют Малой Северной Двиной. Преодолев долгий и извилистый путь, на севере в реку впадает Пинега. Дальше начинаются многочисленные разветвления, состоящие из протоков с множеством островов, ширина долины составляет около 18 км.

После города Новодвинска начинается дельта, площадь которой примерно 900 км^2. Затем река впадает в Двинскую губу Белого моря. Этот путь можно проследить по рисунку 1.

Рисунок 1 – Схема бассейна Северной Двины

Северная Двина, крупнейшая река Европейского Севера России, образуется от слияния рек Сухоны и Юга. Длина реки от места слияния до впадения в Двинский залив Белого моря 750 км. Площадь бассейна 357000 км^2. По величине бассейна Северная Двина занимает пятое место среди рек Европейской части России. Сток Северной Двины (110 км^3 в год) составляет около трети общего речного стока в Баренцево и Белое моря.

Северная Двина - типичная равнинная река с плавным продольным профилем, сравнительно небольшими уклонами (средний уклон около 0,07 %) и широкой долиной, пойма которой достигает 10 км и более.

В весенний период пойма реки затопляется. Извилистое русло реки изобилует островами, осередками и песчаными перекатами, затрудняющими судоходство.

Климатические условия бассейна, определяющие гидрологический режим реки, характеризуются продолжительной и холодной зимой, коротким и прохладным летом и большим количеством осадков. За год в бассейне выпадает в среднем около 500 мм осадков, причем 60 - 70 % их приходится на теплую половину года. Бассейн Северной Двины, находящийся под влиянием влажных воздушных масс, поступающих с запада, относится к зоне избыточного увлажнения. Эти климатические условия определяют сравнительно высокий модуль стока, равный для всего бассейна в среднем 9,7 л/сек с 1 км^2 при коэффициенте стока 0,61.

Гидрологический режим Северной Двины характеризуется высоким весенним половодьем, сравнительно низкой летней меженью с дождевыми паводками и низкими уровнями зимой. Благодаря накоплению в течение длительной зимы осадков в виде снега и интенсивного весеннего таяния объем стока весеннего половодья достигает 50 % годовой величины стока воды. В формировании гидрологического режима Северной Двины существенную роль играет направление течения с юга на север. Это особенно проявляется в весенний и осенний периоды.

Весеннее таяние, начинающееся на юге, в верховьях, обусловливает образование паводка, движение которого совпадает с направлением движения весны, т.е. с юга на север. В этих случаях продвижение паводочной волны сопровождается пополнением ее благодаря таянию снега на нижележащих участках, где оно происходит позднее. Паводочная волна, двигающаяся в сторону еще не вскрывшихся участков реки, способствует освобождению их ото льда. Вместе с тем происходит концентрация на отдельных участках больших масс взломанного льда с образованием временных остановок ледохода - заторов. Они вызывают стеснение живого сечения и резкие и высокие подъемы уровней. Заторные явления характерны для всей реки - от верховьев до устья.

Вскрытие реки происходит на подъеме весеннего половодья, максимум которого наблюдается при ледоходе. Спад весеннего половодья бывает более медленным, чем подъем. Весеннее половодье начинается на Северной Двине во второй половине апреля - начале мая, и расход воды возрастает до 11 - 36 тыс. $м^3/с$.

Весной по реке приносится более половины общего годового стока - снеговые воды занимают в ее питании 51%. В орографическом отношении бассейн представляет собой обширную лесистую, слабо всхолмленную равнину, приподнятую по краям и понижающуюся в северо-западном направлении. Вся поверхность бассейна покрыта мощным слоем ледниковых отложений, подстилаемых песчаниками, мергелями и известняками.

Бассейн находится в зоне тайги, представленной елью, сосной, лиственницей. Залесенность территории бассейна 80 - 85 %.

Значительная площадь бассейна (около 8,5 %) заболочена. Много моховых болот. Вечная мерзлота отсутствует.

1.1 Рельеф

К верхнему течению реки, относят участок от истока до впадения Вычегды, на этом участке образована широкая долина, берега реки в ней высокие и крутые, состоящие из известняковых и песчаных пластов.

Среднее течение, верхняя граница которого – это место впадения Вычегды, а нижняя – устье реки Ваги. Оно так же расположено в широкой долине, береговые возвышенности которой попеременно отличаются друг от друга. Правый берег обычно круче левого, состоит из глинистых утесов, а левый луговой. В некоторых местах происходит наоборот, то есть левый берег крутой, а правый луговой.

Нижнее течение реки отличается наличием холмов и невысоких гор. Возле Архангельска высота берега может быть до двадцати одного метра ввысь, которые состоят из песчано-известковых и глинистых холмов.

В общем случае рельеф равнинный. Равнинный тип местности нарушается из-за наличия хребтов на северо-западной окраине Ветреного Пояса и возвышенностей Тимана в центральной части.

В районе западного Урала, рельеф переходит в горный, вместе с ним меняются климат, растительность, геологическое строение, гидрография и многое другое.

Невысокие плато встречаются как на низинных, так и на возвышенных равнинах, бывают слабоволнистые и всхолмленные.

В районе морских водных объектов преобладают низменности, по мере отдаления от них высота местности увеличивается, причем полосами вдоль главных рек и их наиболее крупных притоков.

Возвышенные равнины приурочены к приводораздельным участкам междуречий. Чем дальше от морского побережья, тем эти участки обширнее. В целом - поверхность Северного края понижается с юга на север, что и определяет общее направление речного стока - к Белому и Баренцеву морям.

1.2 Климат

Огромная протяженность территории и неоднородность рельефа, создают различные климатические условия для различных частей бассейна реки.

Особенности климата определяются малым количеством солнечной энергии зимой, воздействием северных морей, особенно заметным переносом воздушных масс. Для Северной Двины характерна частая смена воздушных масс при прохождении циклонов со стороны Атлантики. С циклонами связана пасмурная погода, нередко с осадками, теплая и часто с оттепелями зимой и прохладная летом. Развитие циклонов наиболее развито зимой и осенью, летом ослабевает. Поступление воздушных масс арктического происхождения в любое время года сопровождается холодными и сухими северо-восточными ветрами, приносящими резкие похолодания. Наиболее часто их вторжения наблюдаются в летнее время. Со стороны Сибири зимой нередко переносится континентальный воздух, принося сухую морозную то погоду. С юга и юго-востока поступают преимущественно континентальные массы воздуха, охлажденные зимой и прогретые летом.

Частая смена воздушных масс придает погоде в течение всего года большую неустойчивость. Влияние морей сильно сказывается на распределении температуры воздуха по территории. Зимой температура воздуха на побережьях морей выше, чем в удалении от моря, а летом - ниже. В глубь материка в направлении с запада на восток ослабевает влияние Атлантики. Следовательно, с севера на юг и с запада на восток нарастает континентальность климата. Совокупность перечисленных факторов обусловливает короткое прохладное лето и длинную холодную зиму с устойчивым снежным покровом, более мягкую в западных районах Северного края и более суровую в восточных. Небольшие местные различия климатических условий связаны с микро и мезо формами рельефа, экспозицией склонов, близостью озер и болот и др. Зима продолжается пять-шесть месяцев на западе территории, шесть – семь месяцев на востоке. Средняя температура воздуха за наиболее холодный

месяц достигает обычно – 20°С. Снежный покров устойчив. Характерны частые метели: зимой преобладают ветры южного, юго-западного направления, средняя скорость которых 3 - 7 м/сек. Осадков зимой выпадает от 110 до 200 мм; наибольшее количество осадков наблюдается в горах и предгорьях Урала, а в пределах равнинной части территории - на наветренных склонах возвышенностей и уступах плато. Лето продолжается три-четыре месяца в юго-западных районах, один- два месяца в северо-восточных. Средняя месячная температура не превышает 16 - 17°, заморозки возможны в любом из летних месяцев. Ветры преимущественно северного и северо-восточного направлений, их скорость 2,5 - 3,5 м/сек., на побережьях морей - 4,5 - 5,5 м/сек. Осадков за летние месяцы выпадает 400 - 500 мм.

1.3 Хозяйственное использование поверхностных вод

Реки Северного края используются в основном для судоходства и сплава леса. Водозабор из рек и озер и использование их энергетического потенциала незначительны.

Рыболовство развито слабо и имеет местное значение, исключая лов семги (благородного лосося).

Регулярное судоходство производится на реках Онеге, Северной Двине, Сухоне, Вологде, Вычегде, Сысоле, Ваге, Емце, Мезени и Печоре. На большом протяжении и почти вне зависимости от водности года оно обеспечивается только на реках Северной Двине, Сухоне, Вычегде и Печоре. На остальных реках судоходство возможно только на наиболее многоводных участках протяжением от нескольких десятков километров (Вологда, Сысола, Емца) до 100 - 200 км (Онега, Вага, Мезень). В очень маловодные годы эти участки в период межени значительно сокращаются или вовсе становятся несудоходными. Общая протяженность водных путей около 4 тыс. км. Судоходство производится в течение 5 - 6 месяцев, с мая по октябрь.

Для поддержания гарантийных габаритов водных путей ежегодно выполняется большой объем землечерпательных работ на лимитирующих перекатах. Годовой объем грузоперевозок около 30 млн. т. Половина из них приходится на лесные грузы (в буксируемых плотах). Пассажиров перевозится 6 - 7 млн. человек, в основном на местных линиях. Основной объем грузоперевозок приходится на реки бассейна Северной Двины.

Сплав леса производится на всех больших и на многих малых реках, расположенных в пределах таежной зоны. На реках Северной Двине, Вычегде, Мезени и Печоре он плотовой, причем плоты транспортируются буксирами. Лес сплавляется в течение всего навигационного периода. На остальных реках сплав молевой и производится в основном в период весеннего половодья. Общее протяжение сплавных путей около 27 тыс. км, в основном они сосредоточены в бассейне Северной Двины.

Ежегодный объем сплавляемой древесины 30 - 40 млн. $м^3$. Осадка плотов обычно не превышает 1,0 - 1,2 м. По Северной Двине провозятся и большегрузные плоты, имеющие осадку 1,5 - 1,8 м и длину 300 - 400 м, объем леса в таком плоту достигает 15-25 тыс. $м^3$. Проводка их в межень через обмелевшие перекаты затруднительна, особенно при ветре. Конечными пунктами сплава являются города Онега, Архангельск, Мезень и Нарьян-Мар, где расположено большинство лесозаводов и предприятий по химической переработке древесины. Однако большая часть сплавляемой древесины выкатывается на берег в местах ее перевалки на железную дорогу еще в верхней части рек бассейнов Северной Двины и Печоры и отправляется вглубь страны в виде круглого леса. Наиболее благоприятными для сплава являются средние уровни на спаде весеннего половодья. При очень высоких половодьях нередки аварии в системе ограждения сплавных путей и разнос древесины по пойме; отмечаются также срывы запаней или преждевременное начало молевого сплава, связанное с разрушением складов леса, расположенных на берегу. При быстром спаде уровней во время маловодных половодий много леса оседает на берегах и мелях, образуя заломы. Водоза-

боры из рек для промышленных и коммунально-бытовых нужд незначительный; проблема водообеспечения существует только для отдельных городов, расположенных на маловодных реках (на реках Воркуте, Вологде и др.), в наиболее маловодные периоды. Столь же незначителен и сброс в реки загрязненных вод, однако из-за недостаточной очистки последних и малой водности рек - водоприемников сточные воды местами вызывают значительное ухудшение качества речных вод. Потенциальная мощность рек Северного края оценивается в 6,6 млн. кВт-ч, а технически возможная годовая выработка энергии около 25 млрд. кВт-ч. Наибольшим энергозапасом обладает р. Печора.

1.4 Описание поста

Пост Звоз расположен возле населенного пункта, расположенного в 0,8 км выше впадения р. Большой Кироксы.

Прилегающая к долине реки местность - равнинная, представляет собой карстовую область с многочисленными воронкообразными и оврагообразными провалами. Река в районе поста прорезает толщу Пермских гипсов, образуя ящикообразную долину, шириной до 1,5 км. Склоны долины крутые, преимущественно обнаженные, высотой 20-25 м, рассечены глубокими оврагами.

Пойма на участке поста правобережная, шириной 50-100 м, в 2 км ниже водпоста переходит в левобережную, шириной до 500 м, занята лугом и кустарником, затопляется при уровне 700 см над нулем графика.

Русло реки слабоизвилистое, песчаное, у левого берега - каменистое, деформирующееся. В 5 км выше водпоста расположен Шепиловскй перекат, в 2 км ниже - перекат Звозский (бывший Медведский). Остров Медведский, находившийся в 2.5 км ниже водпоста, летом 1959 г. уничтожен путем взрывных работ и работы земснаряда. Весной при ледоходе наблюдается

подпор уровня от заторов льда ниже водпоста с подъемом уровня до 5 м над бесподпорным.

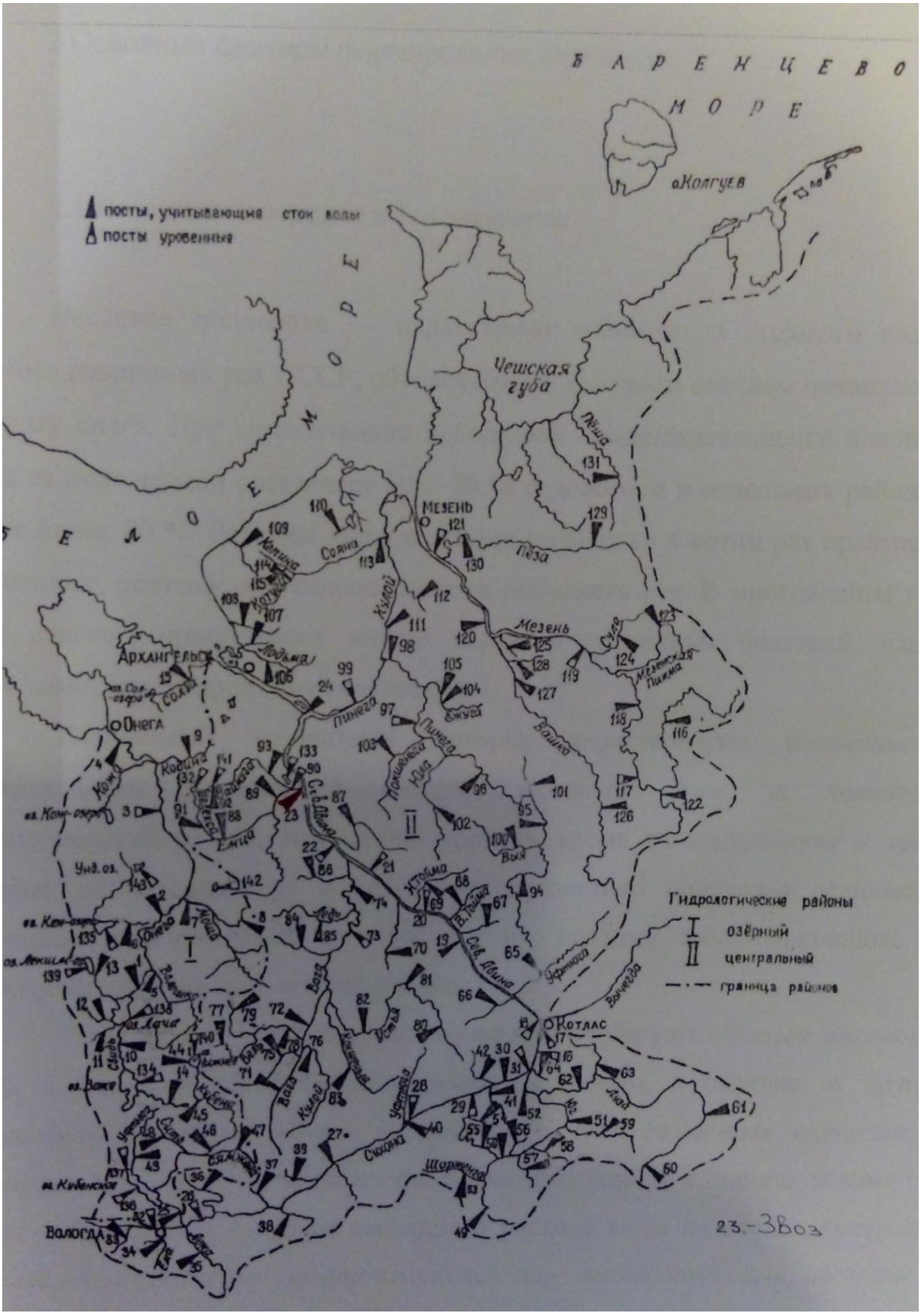


Рисунок 2 – Схема расположения постов

В зимний период в районе водпоста ежегодно наблюдаются большие скопления шуги. Водпост находится на левом берегу оборудован деревянными и металлическими сваями. Отметки водомерному устройству переданы нивелировкой IV класса АУГМС 5.05.1951 г. Длина хода 0,2 км. Отметка нуля поста 5,86 м БС. Верхним уклонным постом является водпост гидроствора № 1, нижним основной водпост. 1 идроствор Ms 1 (паводочный) расположен в 2.5 км выше основного водпоста, гидроствор Ms 2 (меженный) - в створе основного водпоста, гидроствор М» 3 - в 0,6 км выше основного водпоста. Температура воды измеряется в створе водпоста у берега, толщина льда в створе водпоста на середине реки.

Прилегающая к долине реки местность - равнинная, представляет собой карстовую область с многочисленными воронкообразными и оврагообразными провалами. Река в районе поста прорезает толщу Пермских гипсов, образуя ящикообразную долину, шириной до 1,5 км. Склоны долины крутые, преимущественно обнаженные, высотой 20-25 м, рассечены глубокими оврагами. Пойма на участке поста правобережная, шириной 50-100 м, в 2 км ниже водпоста переходит в левобережную, шириной до 500 м, занята лугом и кустарником, затопляется при уровне 700 см над нулем графика. Русло реки слабоизвилистое, песчаное, у левого берега - каменистое, деформирующееся. В 5 км выше водпоста расположен Шепиловскй перекат, в 2 км ниже - перекат Звозский (бывший Медведский). Остров Медведский, находившийся в 2.5 км ниже водпоста, летом 1959 г. уничтожен путем взрывных работ и работы земснаряда. Весной при ледоходе наблюдается подпор уровня от заторов льда ниже водпоста с подъемом уровня до 5 м над бесподпорным. В зимний период в районе водпоста ежегодно наблюдаются большие скопления шуги. Водпост находится на левом берегу оборудован деревянными и металлическими сваями. Отметки водомерному устройству переданы нивелировкой IV класса АУГМС 5.05.1951 г. Длина хода 0,2 км. Отметка нуля поста 5,86 м БС. Верхним уклонным постом является водпост гидроствора

№ 1, нижним основной водпост. 1 идроствор Ms 1 (паводочный) расположен в 2.5 км выше основного водпоста, гидроствор Ms 2 (меженный) - в створе основного водпоста, гидроствор М» 3 - в 0,6 км выше основного водпоста. Температура воды измеряется в створе водпоста у берега, толщина льда в створе водпоста на середине реки.

2 Методология нахождения поставленной цели

Данная работа является попыткой использовать методы машинного обучения для анализа гидрологических данных. Существует межотраслевой стандарт решения задач интеллектуального анализа данных CRISP-DM, который представлен на рисунке 3.

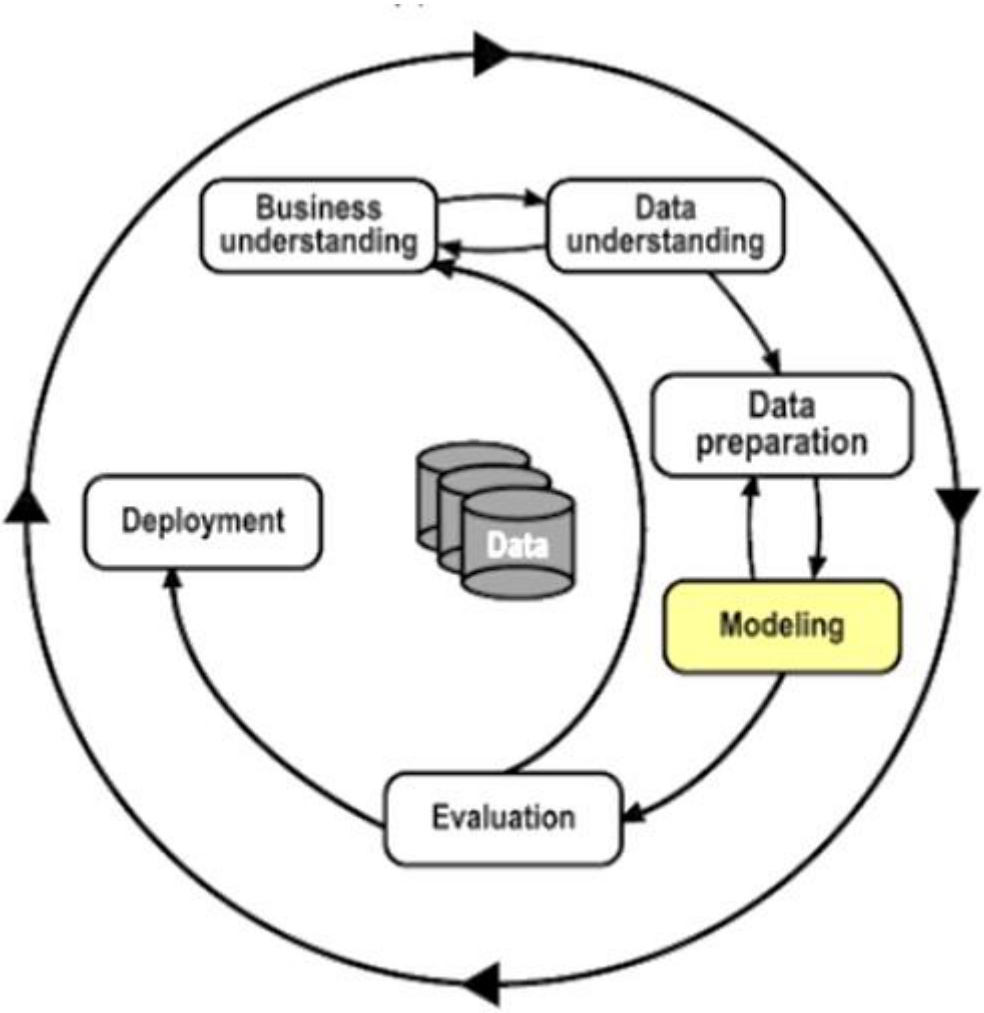


Рисунок 3 – межотраслевой стандарт решения задач интеллектуального анализа данных CRISP-DM

Он предполагает решение задачи в пяти основных шагах, причем в цикле они могут замыкаться и повторяться многократно:

1) Понять предметную область задачи.

2) Выяснить как собирались данные:

а) проверить наличие шумов, пропусков, выбросов

б) вычислить информативность признаков.

3) Подготовить данные для моделирования.

4) Построение предсказательной модели.

5) Использование модели на практике.

Выполнению этих пяти шагов и будет посвящена эта работа с акцентом на четвертый шаг, он является основным и предполагает основную задачу машинного обучения.

К задачам машинного обучения относят задачи предсказаний, принятий решений, поиск сложных закономерностей. Методы решения этих задач могут быть применимы в гидрологии, например, для прогнозирования различных гидрологических характеристик. Чтобы решить эти задачи желательно располагать большим объемом данных.

Зависимости между стоком на начало половодья, максимальными снегозапасами в лесу, суммой температур за зиму, датой ледостава и, например, датой начала весеннего половодья достаточно сложно выявить, поэтому возникает необходимость в поиске алгоритма, позволяющего решить эту задачу.

По сути машинное обучение (МО) решает задачу восстановления функции по точкам. Другими словами, имеется неизвестная функция отображения из множества объектов в множество решений, которая промерена в конечном множестве точек. Значения этой функции в определенном множестве точек образует обучающую выборку. Используя обучающую выборку, состоящую из пар объект-ответ, при помощи методов машинного обучения решается задача восстановления зависимости между этими парами или построение функции, которая бы аппроксимировала вот эту самую зависимость.

Прежде чем приступать к реализации методов, нужно понять, как задаются объекты, что такое ответы, как строить аппроксимирующую функцию, и как оценивать её качество.

После применения методики прогноза, необходимо оценить её качество. В данной работе эта оценка делалась привычным в гидрологии способом.

Для долгосрочных прогнозов допустимая погрешность рассчитывалась по формуле:

$$\delta_{доп} = \pm 0,674\sigma, \quad (1)$$

где σ – среднее квадратическое отклонение прогнозируемого значения элемента от среднего:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}} \quad (2)$$

где y_i –значение прогнозируемой величины;

$\overline{y}$ – среднее значение;

n – количество элементов ряда.

Таблица 1 – Допустимая погрешность прогноза $\delta_{доп}$

Дата начала половодья	Расход на начало половодья	Дата пика половодья	Максимальный расход половодья	Дата конца половодья	Расход конца половодья
6	149	6	2390	9	421

Прогноз будет считаться оправдавшимся, если абсолютная величина его погрешности меньше или равна допустимой.

За меру точности методики прогнозирования была принята средняя квадратическая погрешность проверочных прогнозов, вычисляемая по формуле

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n - m}}, \quad (3)$$

где y_i и y' – соответственно фактическое и предсказанное значения;

n – количество элементов ряда;

m – число степеней свободы, равное числу постоянных в прогностическом уравнении.

Критерий качества рассчитывался по отношению S / σ.

При $n \geq 25$, методика считается эффективной, если $S / \sigma \leq 0{,}80$.

3 Семейства алгоритмов машинного обучения

Существует множество различных алгоритмов машинного обучения для решения задач классификации и регрессии, в этой работе для задач классификации и регрессии были использованы алгоритмы решающих деревьев, которые относятся к логическим методам и метод многослойной нейронной сети, так как они способны обнаруживать сложные структуры и связи между разнородными признаками и в теории должны были справиться с поставленной целью.

Выделяют три вида машинного обучения:

1) контролируемое;

2) неконтролируемое;

3) обучение с подкреплением.

Контролируемое обучение еще называют обучением с учителем, здесь каждому объекту обучающей выборки присваивается некоторый заранее известный ответ.

Неконтролируемое обучение или обучение без учителя происходит по точкам данных без различных меток, ответов. Часто методы этого вида решают задачу кластеризации данных.

Обучение с подкреплением – это такое обучение, при котором алгоритм сам выбирает действие, как реакцию на значение точки обучающей выборки.

Существуют ещё байесовские методы в которых предполагается, что данные в каждой точке статистически независимые.

4 Обработка исходных данных

Набор данных, по которым строится модель, называется обучающей выборкой. Обучающая выборка представлена в виде матрицы. Строки этой матрицы – объекты, столбцы – это соответствующие им признаки. Важно, что каждой строчке соответствует столбец с правильными ответами. В зависимости от задачи ответы могут быть действительными числами, в случае регрессии, для классификации это – номер класса.

В качестве объектов исследования выступают даты и величины расхода воды начала и конца весеннего половодья Северной Двины. Признаки — это функции, которые в соответствие объектам ставят какие-то значения. На содержательном уровне это какие-то способы измерения над объектами, и в зависимости от того, что это за измерения, признаки делятся на следующие типы: самый простой это бинарный признак, это какой-то ответ «да» или «нет» про интересующий нас объект. Чуть сложнее номинальные признаки, они принимают тоже конечное множество значений, но это уже большее число значений. Признаки порядковые — это когда на множестве значений признаков задано некое отношение порядка. И, наверное, самый распространённый способ — это количественные признаки, которые являются какими-то числовыми измерениями над нашими объектами [3].

В гидрологии бывают задачи, в которых эти типы смешаны, и в качестве признаков объекта мы имеем и бинарные, и номинальные, и количественные, и может быть даже ещё какие-то более сложные измерения, но в нашем случае мы имеем только количественные признаки, такие как: суммарный слой стока за половодье (мм), максимальные снегозапасы в лесу, сумма отрицательных среднесуточных температур воздуха за зимний период, дата ледостава и толщина льда на 10 апреля (мм).

Для того чтобы значения дат можно было использовать как количественный признак, пришлось преобразовать их в количество дней от первого января соответствующего года до значения той или иной временной даты.

Итак, у нас есть набор признаков, встает вопрос – все ли они одинаково полезны для того или иного моделирования объекта? Нахождение ответа на этот вопрос – это один из самых важных этапов подготовки данных для моделирования, от него зависит успех дальнейших вычислений, так как присутствие в данных неинформативных признаков приводит к снижению точности многих моделей, особенно линейных, таких как линейная и логистическая регрессия.

4.1 Отбор признаков

Методы предобработки данных в англоязычной литературе называют Feature Selection и Feature Engineering. Для методов Future Engineering нет готовых алгоритмов, их создание требует творческого подхода и экспертных знаний. Зато для Feature Selection есть готовые реализации в свободном доступе. Например, алгоритм случайного леса, который был использован в данной работе, есть в библиотеке Scikit-Learn. Scikit-Learn – это библиотека, используемая для машинного обучения. Она с открытым кодом, который написан на языке программирования Python. Подробнее о методе случайного леса будет изложено в пятой главе.

Расчет информативности признаков при помощи метода случайного леса был реализован методами библиотеки Scikit-Learn следующим образом:

```
from sklearn import metrics
from sklearn.ensemble import ExtraTreesClassifier
model = ExtraTreesClassifier()
model.fit(X, y)
importances=model.feature_
importances_
```

Для того чтобы использовать менее информативные признаки в расчетах, нужно уменьшить размерность данных, при этом учесть информатив-

ность урезаемого признака. Для этого существует ряд методов: метод главных компонент, многомерное шкалирование, метод независимых компонент и другие. Для нашего случая будем использовать метод главных компонент, так как в литературе он упоминается как одни из основных методов понижения размерности данных.

Главные компоненты находят через вычисление собственных векторов и собственных значений ковариационной матрицы исходных данных или через сингулярное разложение данной матрицы. В литературе его так же можно встретить под названием преобразование Кархунена-Лоэва или преобразование Хотеллинга. Математическое описание метода можно найти здесь [4].

4.1.1 Масштабирование данных

Некоторые алгоритмы машинного обучения сильно чувствительны к калиброванию данных, например, нейронные сети. Поэтому перед их использованием обычно делается нормализация или так называемая стандартизация. Нормализация предполагает преобразование значений признаков так, чтобы они лежали в диапазоне от нуля до единицы. По формуле

$$k_n = \frac{k_i - k_{\min}}{k_{\max} - k_{\min}}, \tag{4}$$

Стандартизация же подразумевает такую предобработку данных, после которой каждый признак имеет нулевое среднее единичную дисперсию.

$$k_{st} = \left(\frac{k_i}{\overline{k}} - 1 \right) \Big/ Cv \tag{5}$$

В библиотеке Scikit-Learn, разработанной для реализации на языке программирования Python, есть готовые функции, позволяющие провести нормализацию данных. Реализация выглядит следующим образом:

```
from sklearn import preprocessing
#Нормализация признаков
normalized_X = preprocessing.normalize(X)
#Стандартизация признаков
standardized_X = preprocessing.scale(X)
```

4.1.2 Кластерный анализ

Для успешного обучения нужно, чтобы одному ответу было соотнесено несколько образов объекта. Так как у нас ответы – это различные величины за каждый год, нужно эти величины объединить в группы или кластеры. Предсказание будет выполняться по среднему значению в кластере, другими словами, при успешном прогнозировании принадлежности проверочного образа к тому или иному кластеру, в качестве прогноза будет взято среднее значение ответов образов, входящих в этот кластер.

Кластеризацию относят к классу задач обучения без учителя.

Кластерный анализ или, по-другому, анализ сегментации, таксономический анализ, разделяет данные на непересекающиеся подмножества (группы). Эти группы должны быть сформированы так, чтобы объекты в одной были схожи, а в разных – различны. Есть статистические методы разделения на группы (кластеры) и методы машинного обучения, и те, и другие используют меры сходства (меры расстояния) для создания кластеров. Например, можно использовать евклидово расстояние, взвешенное евклидово расстояние, метрику Хемминга, Минковского или другую. Для разнородных данных рекомендуется провести масштабирование.

Постановка задачи:

1) Есть множество объектов X и множество номеров групп Y.

2) Задана функция расстояния между объектами $\rho(x, x')$.

2) Имеется конечная обучающая выборка объектов

$$X^m = \{x_1, \cdots, x_m\} \subset X \qquad (6)$$

Требуется разделить выборку на не пересекающиеся подмножества (группы), так, чтобы каждая группа состояла из объектов, близких по метрике ρ, а объекты разных групп существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер y_i.

Для решения поставленной задачи, необходимо разработать алгоритм решения или воспользоваться существующими.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер группы $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число групп, с точки зрения того или иного критерия качества.

Методы кластеризации:

1) Графовые алгоритмы

2) Статистические алгоритмы

а) Алгоритм k-средних

б) EM – алгоритм

3) Алгоритм ФОРЕЛЬ

4) Иерархическая кластеризация или таксономия

5) Нейронная сеть Кохонена

6) Ансамбль кластеризаторов

Подробно и простым языком об этих методах написано в книге [5].

В данной работе использовался метод k-средних, так как он прост в реализации и в литературе упоминается как метод, который дает неплохие

результаты и нейронная сеть Кохонена, потому что она дает наглядное представление о формируемых кластерах, за счет чего проще подобрать оптимальное число разбиений.

4.1.2.1 Метод k-средних

Алгоритм кластеризации методом k-средних:

Шаг 1: нужно сформировать начальное приближение центров всех кластеров $y \in Y : \mu_y$ – наиболее удаленные друг от друга объекты выборки;

Шаг 2: отнести каждый объект к ближайшему центру:

$$y_i := \arg\min \rho(x_i, \mu_y),\ i = 1,\ldots,\ell, \tag{7}$$

Шаг 3: вычислить новое положение центров:

$$\mu_{yj} := \frac{\sum_{i=1}^{\ell}[y_i = y] f_j(x_i)}{\sum_{i=1}^{\ell}[y_i = y]},\ y \in Y,\ j = 1,\ldots,n. \tag{8}$$

Шаг 4: Повторять шаг 2 и шаг 3 до тех пор, пока y_i не перестанут меняться.

Для использования данного алгоритма важно правильно выбрать начальные приближения центров и заранее задать количество кластеров. Центры рекомендуется выделять, последовательно выбирая наиболее удаленные точки. Например, две точки выбираем по наибольшему значению всех попарных расстояний, затем ищем такую точку, расстояние от которой до выделенных ранее наибольшее. Эту процедуру приходится выполнять несколько раз, сравнивая критерии качества, в данной работе таким критерием было отклонение элементов одного кластера от его среднего значения. Для выбора первых двух точек можно построить вспомогательные графики

зависимости одной любой переменной от другой. Для выбора количества кластеров k использовался метод Кохонена. На практике, результаты, полученные методом k-средних, оказались не на много лучше результатов, полученных методом карт Кохонена, поэтому этим методом были рассчитаны только характеристики начала половодья, они приведены в приложении А.

4.1.2.2 Самоорганизующиеся карты Кохонена

Для наглядного представления расположения центров кластеров, расстояния между ними, количества элементов в них, хорошо подходят карты Кохонена (self-organizing maps, SOM).

Построить карту, значит спроецировать объекты выборки на плоскость, в данном случае, на множество узлов прямоугольной сетки заранее заданного размера $M \times N$. Каждый узел занимает нейрон Кохонена с вектором весов $w_{mn} \in \mathbb{R}^n, m = 1,...,M, n = 1,...,N.$ Про устройство и принципы работы нейронов будет изложено в главе 6.

Алгоритм обучения карты Кохонена методом стохастического градиента:

Вход:

1) X^ℓ – обучающая выборка

2) η–темп обучения

Выход:

Векторы синаптических весов $w_{mn}, m = 1,...,M, n = 1,...,N$

Шаг 1: инициализировать веса:

$$w_{mn} := random\left(-\frac{1}{2MN}\right), \left(\frac{1}{2MN}\right), \qquad (9)$$

Шаг 2: выбрать объект x_i из X^ℓ случайным образом

Шаг 3: WTA: вычислить координаты узла, в который проецируется объект x_i :

$$(m_i, n_i) := a(x_i) \equiv \arg\min_{(m,n)\in Y} \rho(x_i, w_{mn}), \tag{10}$$

Шаг 4**:** для всех $(m,n) \in Y$, достаточно близких к (m_i, n_i)

WTM: сделать шаг градиентного спуска:

$$w_{mn} := w_{mn} + \eta(x_i - w_{mn})K(r((m_i, n_i),(m,n))), \tag{11}$$

Шаг 5: повторять шаги 2-4, пока размещение всех объектов в узлах не стабилизируется.

где WTA – это правило жесткой кластеризации, про него более подробно можно прочитать здесь [5], $a(x)$ выдает пару индексов $(m,n) \in Y$, которые показывают в какой узел сетки проецируется объект *x*. Карта отражает кластерную структуру выборки если близкие объекты попадают в близкие узлы сетки.

WTM – это правило мягкой конкуренции, оно схоже с WTA, только вместо метрики $\rho(x, x')$, определённой на пространстве объектов, используется евклидова метрика на множестве узлов сетки Y :

$$r((m_i, n_i),(m,n)) = \sqrt{(m-m_i)^2 + (n-n_i)^2}. \tag{12}$$

$K(\rho)$ – это ядро сглаживания:

$K(\rho) = \exp(-\beta\rho^2)$.

Параметр β задает степень сглаженности карты: чем меньше β, тем мягче конкуренция нейронов, тем самым границы кластеров выглядят более сглаженными. Имеет смысл увеличивать значение параметра β с каждым шагом, чтобы сеть Кохонена сначала обучилась кластерной структуре в общих чертах, а затем сконцентрировалась на деталях.

Алгоритм останавливается, когда проекции всех, или хотя бы большинства, объектов выборки $(m_i, n_i) = a(x_i)$ перестанут меняться от итерации к итерации.

В результате расчетов получаем матрицы «объект – номер кластера» и несколько отображений структуры кластеров, которые можно посмотреть в приложении А.

Отображение может быть в виде карты расстояний между ними, как на рисунке 4, расстояния выражены через весовые коэффициенты нейронов, расположенных в узлах сетки. Чем больше расстояние, тем темнее цвет.

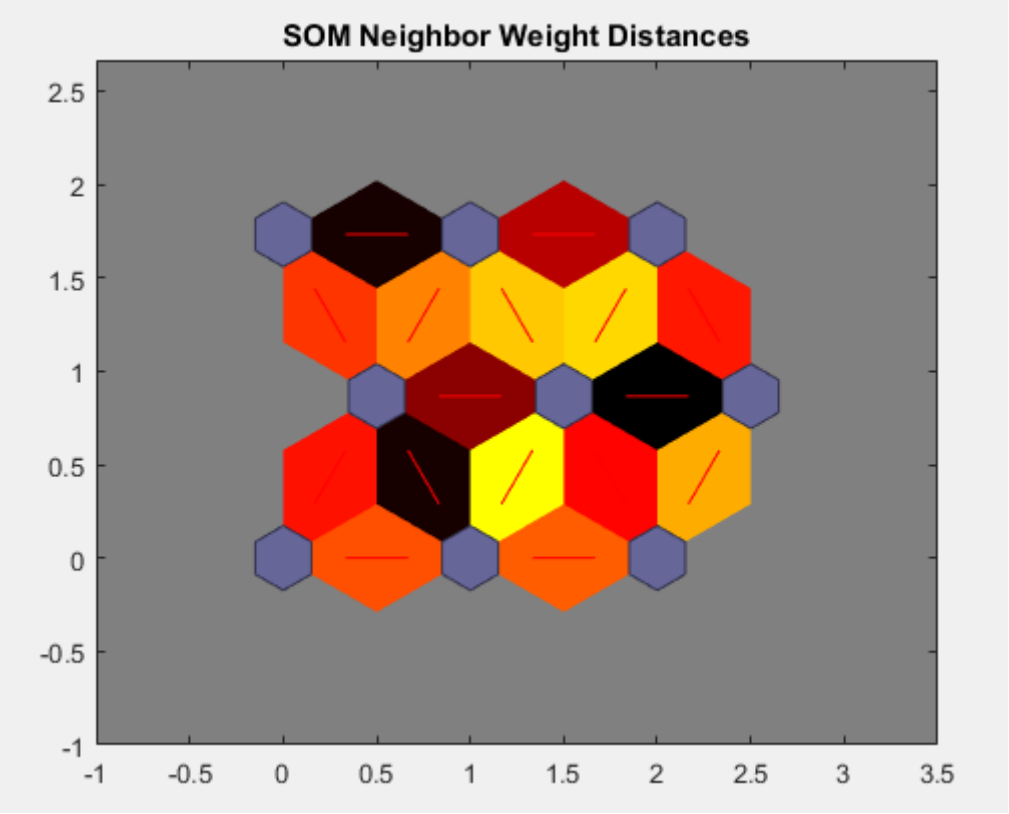


Рисунок 4 – Карта расстояний для даты начала половодья (по нормированным данным)

Так же получаем карту количества объектов в каждом кластере, как на рисунке 5.

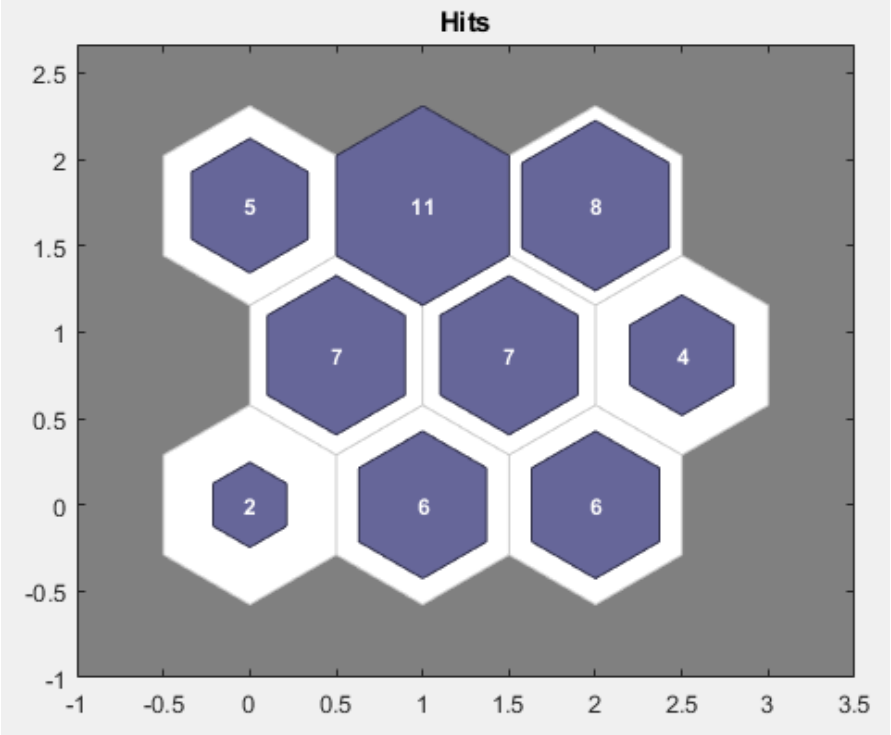


Рисунок 5 – Количество объектов кластерного анализа для даты начала половодья (по нормированным данным)

Далее, на рисунке 6 представлена карта расположения центров кластеров.

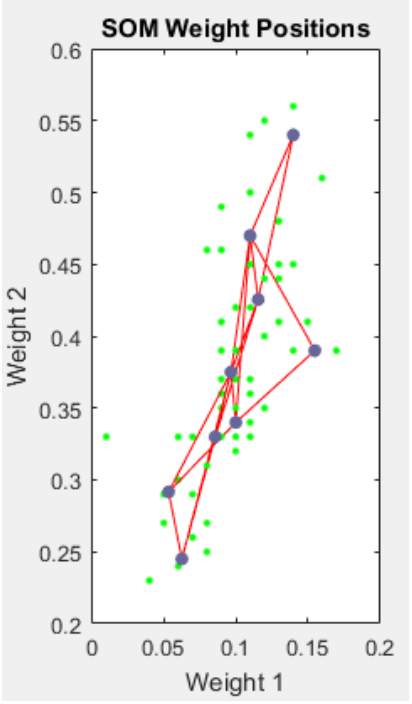


Рисунок 6 – Центры кластеров для дат начала половодья по нормированным данным

5 Решающие деревья

Решающие деревья – это класс логических методов машинного обучения.

Возникновение этих методов связано с попыткой воссоздать способ мышления человека при принятии того или иного решения, проще говоря человеческую логику. Изначально эти методы предназначены для задач классификации, но также они неплохо справляются с задачей регрессии. Основная идея состоит в том, что объединяется определенное количество простых решающих правил, благодаря этому итоговый алгоритм является интерпретируемым.

Постановка задачи: имеется обучающая выборка в виде матрицы «объекты-признаки» и вектора ответов. Требуется найти алгоритм, способный классифицировать новые объекты в виде последовательности принимаемых решений. Математически это можно записать следующим образом.

Дано:

1) векторы $x_i = (x_i^1, ..., x_i^n)$ – объекты обучающей выборки,

2) $y_i = y(x_i), i = 1, ..., \ell$ – классификации или ответы учителя:

$$X \begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \overset{y^*}{\rightarrow} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix} \qquad (13)$$

Найти: функцию $a(x)$, способную классифицировать объекты тестовой выборки $\tilde{x}_i = (\tilde{x}_i^1, ..., \tilde{x}_i^n), i = 1, ..., k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^n \\ \dots & \dots & \dots \\ \tilde{x}_k^1 & \dots & \tilde{x}_k^n \end{pmatrix} \overset{a?}{\rightarrow} \begin{pmatrix} a(\tilde{x}_1) \\ \dots \\ a(\tilde{x}_k) \end{pmatrix} \quad (14)$$

В общем случае решают задачу обучение с учителем. Задача классификации сводится к восстановлению зависимости y: X→Y, |Y| <∞ по точкам обучающей выборки (x_i, y_i), i = 1,…, *l*.

Одна из особенностей решающих деревьев заключается в том, что они позволяют получать важности всех используемых признаков. Важность признака можно оценить на основе того, как сильно улучшился критерий качества, благодаря использованию этого признака в вершинах дерева. Критерий качества еще называют критерием информативности, на его основе принимается то или иное решение на каждом шаге любого выбранного алгоритма. Подробнее об этом будет ниже по тексту.

5.1 Построение деревьев

Для того чтобы понять, как устроены решающие деревья для начала следует рассмотреть алгоритм построения самого простого дерева, то есть бинарного.

Определение бинарного дерева следующее — это ациклический граф, в котором есть два типа вершин: либо вершина соединена с двумя дочерними вершинами, либо она не соединена ни с одной вершиной, и тогда она называется листовой (терминальной) вершиной [6].

Алгоритм бинарного решающего дерева, как следует из названия, представляет собой дерево, в котором каждой вершине сопоставлено некоторое логическое правило. В листьях этого дерева записаны числа-предсказания. Чтобы получить ответ, нужно, двигаясь вверх от корня, делать переходы либо в левое, либо в правое поддерево в зависимости от того, выполняется правило из текущей вершины или нет. Дополнительная информа-

ция в каждой внутренней $V_{внутр}$ и в каждой листовой $V_{лист}$ вершинах связаны.

Алгоритм построения бинарного дерева для классификации $a(x)$:

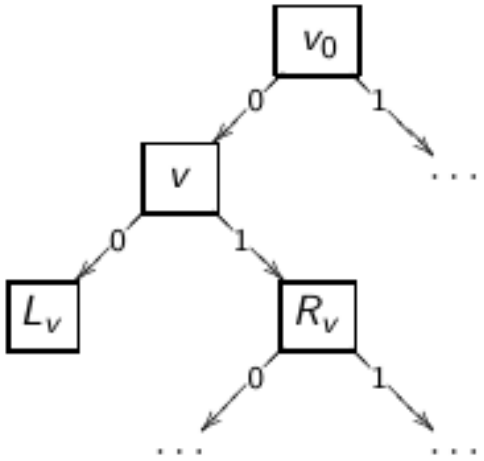
1) $\forall v \in V_{внутр} \rightarrow предикат\ \beta_v : X \rightarrow \{0,1\}, \beta_v \in \mathrm{B},$

2) $\forall v \in V_{лист} \rightarrow имя \quad класса\ c_v \in Y,$

где B – множество бинарных признаков или предикатов

Принятие решений выглядит в виде следующего алгоритма:

```
1: v := v0;
2: пока v ∈ Vвнутр
3:    если βv(x) = 1 то
4:       переход вправо: v := Rv;
5:    иначе
6:       переход влево: v := Lv;
7: вернуть cv.
```



При построении решающего дерева проверяется условие останова и если оно выполнилось, то прекращается рекурсия и вершину, на которой остановились объявляем листом.

Решающее дерево делит все пространство на не пересекающиеся области. Правила называются логическими закономерностями, и характеризуются тем, что каждое такое правило выделяет какую-то область в пространстве объектов, в которой находятся объекты либо только одного какого-то класса, либо преимущественно объекты одного класса.

Каждому листу в соответствие ставится ответ. В задаче классификации – это класс, к которому относится больше всего объектов в листе или вектор вероятностей (вероятность класса может быть равна доле его объектов в

листе). Для регрессии это среднее значение, медиана или другая функция от целевых переменных объектов в листе. Выбор конкретной функции зависит от функционала качества в исходной задаче, обычно, это среднеквадратичное отклонение.

После построения дерева можно провести его «стрижку» – удаление некоторых вершин для понижения сложности и повышения обобщающей способности. Существует несколько подходов к стрижке, о которых можно узнать из лекций Соколова [6].

Таким образом, конкретный метод построения решающего дерева определяется:

1) видом предикатов в вершинах

2) функционалом качества Q (X, j, t)

3) критерием останова

4) методом обработки пропущенных значений

5) методом стрижки.

Часто используют функционал качества вида:

$$Q(R_{m,j,s}) = H(R_m) - \frac{|R_\ell|}{|R_m|} H(R_\ell) - \frac{|R_r|}{|R_m|} H(R_r) \qquad (15)$$

где R_m – это множество объектов в вершине на данном шаге

R_ℓ и R_r – объекты, при заданном предикате, попавшие соответственно в левое и правое поддерево

$H(R)$ критерий информативности.

Критерий информативности дает оценку качества распределения целевой переменной среди объектов множества R. Если разнообразие целевой переменной небольшое, то критерий информативности должен быть меньше, чем, в ином случае. Следовательно, чем меньше его значение, тем объекты

ближе к листовой вершине. Функционал качества при этом должен быть как можно выше.

В общем случае, формула критерия информативности выглядит так:

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c), \qquad 16)$$

где $L(y_i, c)$ некоторая функция потерь.

Таким образом, критерий информативности зависит от вида функции потерь. Функции потерь различны для задач классификации и регрессии. Так же есть критерии, при помощи которых можно получить вероятностную характеристику того или иного ответа, такие как критерий Джини или энтропийный. Чаще всего используют критерий Джини. О нём и других можно прочитать здесь [6].

Критерии останова можно задавать самостоятельно, обычно задаются следующие критерии:

Обычно критериями останова могут стать:

1) ограничение по глубине дерева;

2) ограничение по количеству объектов в листе;

3) в случае, если все объекты в листе относятся к одному классу;

4) ограничение количества листьев в деревьях.

При грамотном подборе приведенных выше критериев и их параметров можно существенно повлиять на качество работы решающего дерева. Этот подбор является трудозатратным и требует проведения кросс-валидации.

Кросс-валидация, она же перекрестная проверка – это метод оценки аналитической модели и её поведения на независимых данных. Данные модели разбиваются на k частей. Обучение модели производят на k−1 частях, а

оставшаяся часть данных используется для проверки. Процедура повторяется k раз. В результате каждая из k частей данных используется для тестирования. В итоге получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

5.2 Градиентный бустинг

Метод градиентного бустинга объединяет в себе различные алгоритмы так, чтобы последующие модели исправляли ошибки предыдущих. Этот метод был предложен Фридманом. Он применим для любых дифференцируемых функций потерь и является одним из наиболее мощных и универсальных на сегодняшний день.

Постановка задачи: имеется обучающая выборка (пары объект – ответ), нужно построить линейную композицию базовых алгоритмов в виде суперпозиции. Количество базовых алгоритмов T задается заранее.

Линейной композицией базовых алгоритмов

$$\alpha_t(x) = C(b(x)), t = 1,...,T, \tag{17}$$

называется суперпозиция функции

$$a(x) = C\left(\sum_{t=1}^{T} \alpha_t b_t(x)\right), \tag{18}$$

где $C: \square \to Y$ – решающее правило;

bt – базовый алгоритм;

α_t –вес базового алгоритма, $\alpha_t \geq 0$.

На выходе получаем линейную комбинацию базовых алгоритмов *bt* с весами *αt*, и при этом ещё применяя решающее правило. Решающее правило

нужно для задачи классификации, для задачи восстановления регрессии решающие правила не нужны, предполагается что *C(b)* – это просто *b*.

Рассмотрим отдельно базовый алгоритм $b(x)$, договоримся что он последний в композиции и будем определять при нём коэффициент, считая, что предыдущие алгоритмы построены и их коэффициенты известны. Наша задача – определить базовый алгоритм *bt,* который будет вычислять очередной вектор u, который содержит ответы всей композиции на всех объектах обучающей выборки. Иными словами, у нас есть вектор u_{T-1}, состоящий из ответов предыдущей композиции, нам нужно построить следующее приближение вектора u_T, в котором каждый i-тый элемент содержит ответ композиции на i-том объекте, в добавок нужно добавить ещё один базовый алгоритм. Математически это выглядит так:

$$b(x) = \sum_{t=1}^{T} \alpha_t b_t(x), \quad x \in X, \quad \alpha_t \in \square \tag{19}$$

Теперь нужно оптимизировать функционал качества с произвольной функцией потерь $\mathcal{L}(b, y)$:

$$Q(\alpha, b) = \sum_{i=1}^{\ell} \mathcal{L}\left(\underbrace{\underbrace{\sum_{t=1}^{T-1} \alpha_t b_t(x_i)}_{u_{T-1,i}} + \alpha b(x_i)}_{u_{T,i}}, y_i \right) \to \min_{\alpha, b} \tag{20}$$

где α – оценка, выданная композицией

b – правильный ответ

$u_{T-1} = (u_{T-1,i})_{i=1}^{\ell}$ – текущее приближение вектора u

$u_T = (u_{T,i})_{i=1}^{\ell}$ – следующее приближение вектора u

Ищем вектор $u = (b(x_i))_{i=1}^{\ell}$ из R^{ℓ}, который приводит к минимуму $Q(\alpha, b)$.

Нужно минимизировать функционал *Q*, который зависит от вектора ответов базового алгоритма на всех объектах обучающей выборки u. Для предположения, что u может быть произвольным, можно использовать градиентный метод оптимизации. Он как раз заключается в том, что фиксируется начальная точка u_0 и затем делаем градиентные шаги

$$Q(u) \to \min, u \in \square^{\ell}: \tag{21}$$

$u_0 :=$ начальное приближение

$$u_{T,i} := u_{T-1,i} - \alpha g_i, \quad i = 1,...,\ell \tag{22}$$

где $g_i = \mathcal{L}'(u_{T-1,i}, y_i)$ – компоненты вектора градиента

α – градиентный шаг.

Добавление базового алгоритма b_T :

$$u_{T,i} := u_{T-1,i} + \alpha b_T(x_i), \quad i = 1,...,\ell \tag{23}$$

Будем искать базовый алгоритм b_T , чтобы вектор $(b_T(x_i))_{i=1}^{\ell}$ приближал вектор антиградиента $(-g_i)_{i=1}^{\ell}$:

$$b_T := \arg\max_b \sum_{i=1}^{\ell} (b(x_i) + g_i)^2 \tag{24}$$

В заключении еще раз кратко опишем алгоритм построения градиентного бустинга. Последовательно строятся базовые алгоритмы, находим і-ый алгоритм так, чтобы он приближал вектор градиента, то есть берем произ-

водную от функции потерь. Затем определяем α каждого алгоритма, это задача одномерной оптимизации. На третьем шаге обновляем значение зиции на объектах выборки.

Алгоритм построения градиентного бустинга:

Вход:

1) обучающая выборка X^{ℓ}

2) параметр T.

Выход: базовые алгоритмы и их веса $\alpha_t b_t, t = 1,...,T$.

Шаг 1: инициализация: $u_i := 0, i = 1,...,\ell;$

Шаг 2: для всех $t = 1,...,T$ найти базовый алгоритм, приближающий градиент:

$$u_i := u_i + \alpha_t b_t(x_i); i = 1,...,\ell. \quad (25)$$

Шаг 3: решить задачу одномерной минимизации:

$$\alpha_t := \arg\min \sum_{i=1}^{\ell} \mathrm{L}(u_i + \alpha b_t(x_i), y_i); \quad (26)$$

Шаг 4: обновить значения композиции на объектах выборки:

$$u_i := u_i + \alpha_t b_t(x_i); i = 1,...,\ell. \quad (27)$$

5.2.1 Расчет характеристик весеннего половодья градиентным бустингом

В результате расчетов были получены значения дат и величин расхода для начала, пика и окончания весеннего половодья. Решались задачи классификации и регрессии. Так же рассчитана информативность признаков модели. Некоторые признаки несут в себе небольшую информативность, поэтому в расчетах эти признаки можно не учитывать, но так как у нас и так мало признаков, они учитывались, но только с меньшим весом. Уменьшение размерности признаков проводилось при помощи метода главных компонент. Об этом упоминалось в главе 4.

В таблицах число главных компонент обозначено как «ГК», серым цветом выделены неправильные ответы значений принадлежности объекта к кластеру.

На рисунках представлены результаты расчета в относительных единицах, как разница между значением полученной даты и даты первого января соответствующего года.

Прогноз даты начала половодья, решая задачу классификации:

Расчет проведен для исходных, нормированных и стандартизированных данных, для выявления влияния преобразования данных на результаты, а также с учетом преобразования данных методом главных компонент.

Ответы прогнозов в виде средних значений кластеров приведены в таблице 2.

Таблица 2 – Ответы прогноза в задаче классификации

Номер кластера	Среднее значение м-д k-средних, k=16	Среднее значение м-д k-средних, k=9 (по нормированным данным)	Среднее значение м-д k-средних, k=9 (по стандартизированным данным)
1	99	92	113
2	103	100	115
3	95	111	111
4	105	100	111
5	117	104	105
6	104	112	97
7	102	108	113
8	111	109	102
9	99	109	94
10	108		
11	109		
12	114		
13	116		
14	102		
15	109		
16	96		

Прогноз даты начала половодья по исходным данным:

Таблица 3 – Информативность признаков для прогноза даты начала половодья

Все признаки	0,090	0,100	0,098	0,117
ГК=3	0,134	0,194	0,145	
ГК=2	0,25	0,26		

Таблица 4 – Прогноз значений даты начала половодья

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	6	6,00	6	6
1951	6	4,00	6	6
1953	7	11,00	11	11
1956	12	12,00	12	12
1967	4	4,00	6	6
1968	15	15,00	15	15
1975	1	1,00	1	1
1978	7	10,00	7	7
1980	6	15,00	10	12

Если посмотреть на ответы кластеров, приведенные в таблице 2, видно небольшую разницу значений между правильными ответами и спрогнозированными. Поэтому, можно сказать, что классификация выдала неплохой результат.

При переходе от значений кластеров к данным получена таблица 5.

Таблица 5 – Прогноз значений даты начала половодья

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	1 май	14 апр	14 апр	14 апр
1951	2 апр	16 апр	14 апр	14 апр
1953	8 апр	20 апр	20 апр	20 апр
1956	29 апр	24 апр	24 апр	24 апр
1967	10 апр	16 апр	14 апр	14 апр
1968	14 апр	18 апр	18 апр	18 апр
1975	9 апр	10 апр	10 апр	10 апр
1978	10 апр	19 апр	13 апр	13 апр
1980	24 апр	18 апр	18 апр	24 апр

S / σ		0,78	0,71	0,68

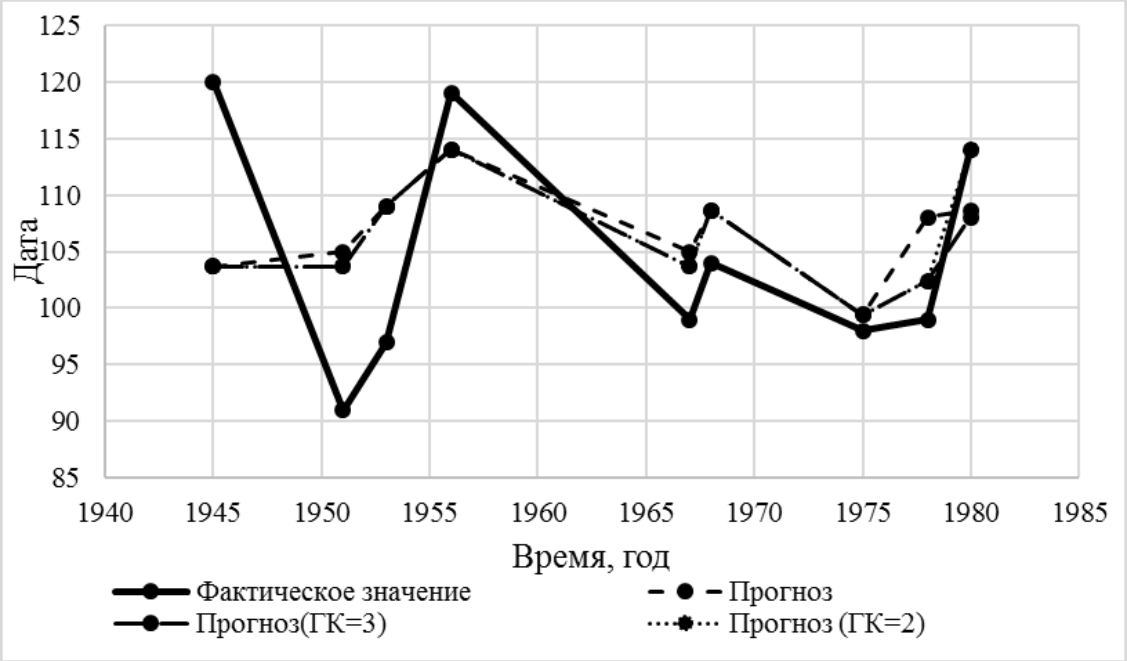


Рисунок 7 – Прогноз значений даты начала половодья

Заметим, что сжатие данных, незначительно повлияло на результаты, но всё же лучшие значения при наименьшем количестве признаков.

Прогноз значения даты начала половодья для нормированных данных:

Таблица 6 – Информативность признаков для прогноза даты начала половодья по нормированным данным

Все признаки	0,07	0,23	0,12	0,03
ГК=3	0,39	0,25	0,35	
ГК=2	0,59	0,41		

Таблица 7 – Прогноз значений даты начала половодья по нормированным данным

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	2	2	2	2
1951	2	8	2	2
1953	5	5	5	5
1956	6	6	6	6
1967	3	2	2	2
1968	9	9	8	8
1975	4	4	4	4

1978	5	5	5	5
1980	5	5	5	5

Таблица 8 – Прогноз значений даты начала половодья по нормированным данным

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	1 май	11 апр	11 апр	11 апр
1951	2 апр	20 апр	11 апр	11 апр
1953	8 апр	15 апр	15 апр	15 апр
1956	29 апр	22 апр	22 апр	22 апр
1967	10 апр	11 апр	11 апр	11 апр
1968	14 апр	18 апр	19 апр	19 апр
1975	9 апр	10 апр	10 апр	10 апр
1978	10 апр	15 апр	15 апр	15 апр
1980	24 апр	14 апр	14 апр	14 апр
S / σ		0,85	0,73	0,72

Из таблицы 8 видно, что критерий качества для прогноза, использующего все признаки, больше 0,80, значит он не оправдался. В таблице 6 значение четвертого признака, которое соответствует значению даты ледостава незначительное, можно сделать вывод что шум, создаваемый этим признаком, снизил качество прогноза. Это говорит о том, что сжатие информации может улучшить качество прогноза. Поэтому при прогнозировании стоит предварительно ознакомиться с информативностью признаков, составляющих входной вектор.

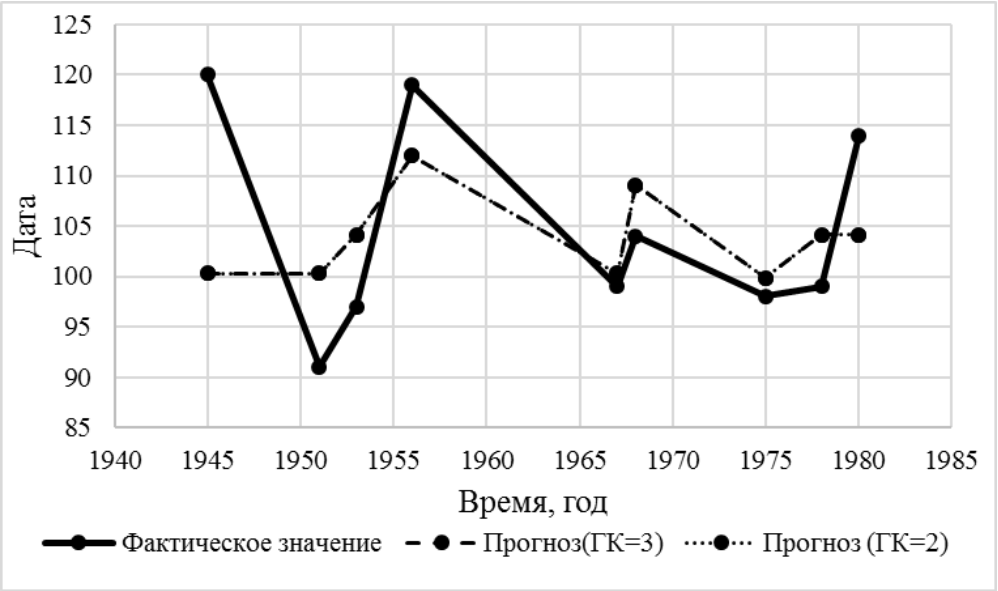


Рисунок 8 – Прогноз значений даты начала половодья по нормированным данным

Прогноз значений даты начала половодья по стандартизированным значениям:

Таблица 9 – Информативность признаков для прогноза даты начала половодья по стандартизированным значениям

Все признаки	0,14	0,10	0,13	0,16
ГК=3	0,35	0,38	0,28	
ГК=2	0,48	0,52		

Таблица 10 – Прогноз значений даты начала половодья по стандартизированным данным

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	4	6	5	6
1951	6	6	5	4
1953	8	7	6	8
1956	2	2	2	2
1967	6	6	6	4
1968	8	8	3	6
1975	6	6	6	6
1978	9	5	6	5
1980	4	4	5	4

Таблица 11 – Прогноз значений даты начала половодья по стандартизированным данным

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	1 май	7 апр	16 апр	7 апр
1951	2 апр	7 апр	16 апр	22 апр
1953	8 апр	24 апр	7 апр	13 апр
1956	29 апр	24 апр	24 апр	24 апр
1967	10 апр	7 апр	7 апр	22 апр
1968	14 апр	12 апр	21 апр	6 апр
1975	9 апр	7 апр	7 апр	7 апр
1978	10 апр	16 апр	7 апр	16 апр
1980	24 апр	21 апр	15 апр	21 апр
S / σ		0,82	0,65	0,94

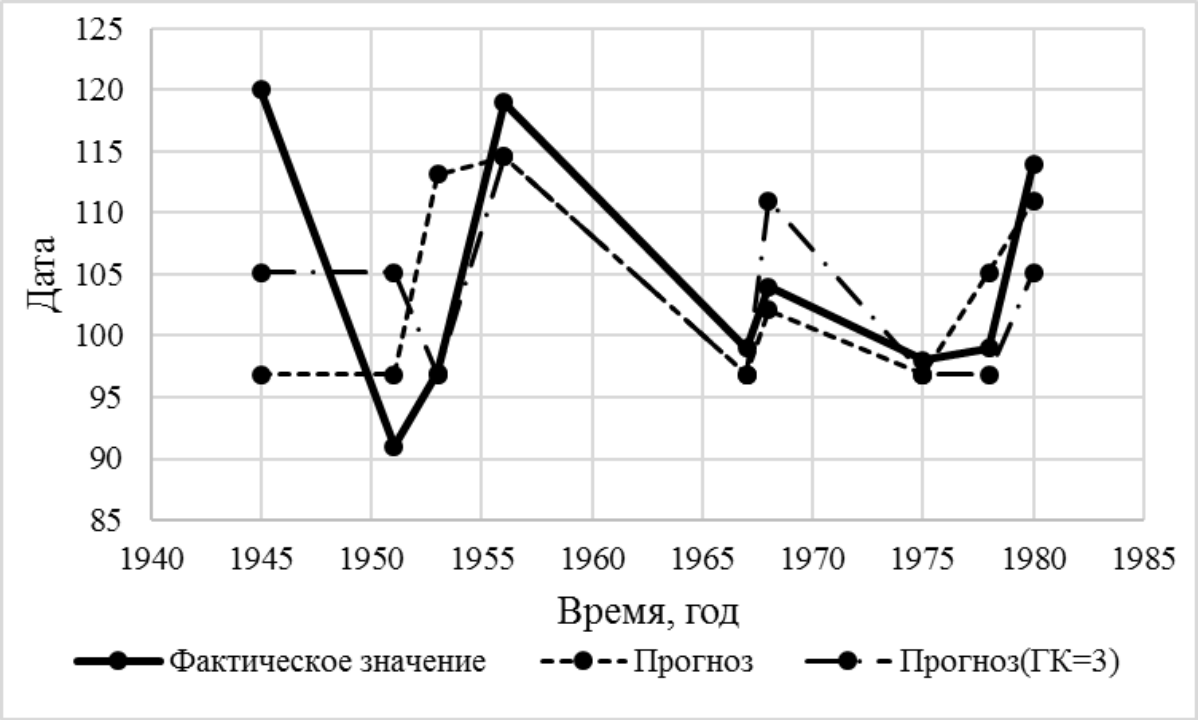


Рисунок 9 – Прогноз значений даты начала половодья по стандартизированным значениям.

На примере прогнозов даты начала половодья, мы видим, что с задачей классификации метод справляется неплохо, ошибки связаны в основном с кластерным анализом, там, где классификатор неправильно классифицировал объекты, наблюдаются близкие ответы кластеров либо небольшое коли-

чество примеров обучающей выборки. Посмотреть ошибки кластерного анализа и число объектов в кластерах можно в приложении А.

Дальнейшее использование классификатора для анализа других характеристик не имеет смысла из-за высокой погрешности кластерного анализа, поэтому для прогноза методом градиентного бустинга остальных характеристик решалась задача регрессии.

Прогноз расхода начала половодья:

Таблица 12 – Информативность признаков для прогноза расхода начала половодья

Все признаки	0,26	0,18	0,36	0,20
ГК=3	0,33	0,30	0,37	
ГК=2	0,54	0,46		

Таблица 13 – Прогноз значений расхода начала половодья

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	616	668,1	596,0	644,7
1951	528	711,2	653,0	683,6
1953	728	780,7	823,7	809,9
1956	572	689,7	565,6	642,7
1967	725	738,9	715,2	617,0
1968	1080	1029,4	1020,4	652,4
1975	968	915,8	881,4	906,5
1978	976	927,4	879,1	812,0
1980	805	853,8	744,2	674,6
S / σ		0,26	0,22	0,53

Как видно из таблицы, качество методики достаточно высокое, лучшее для прогноза по трем признакам.

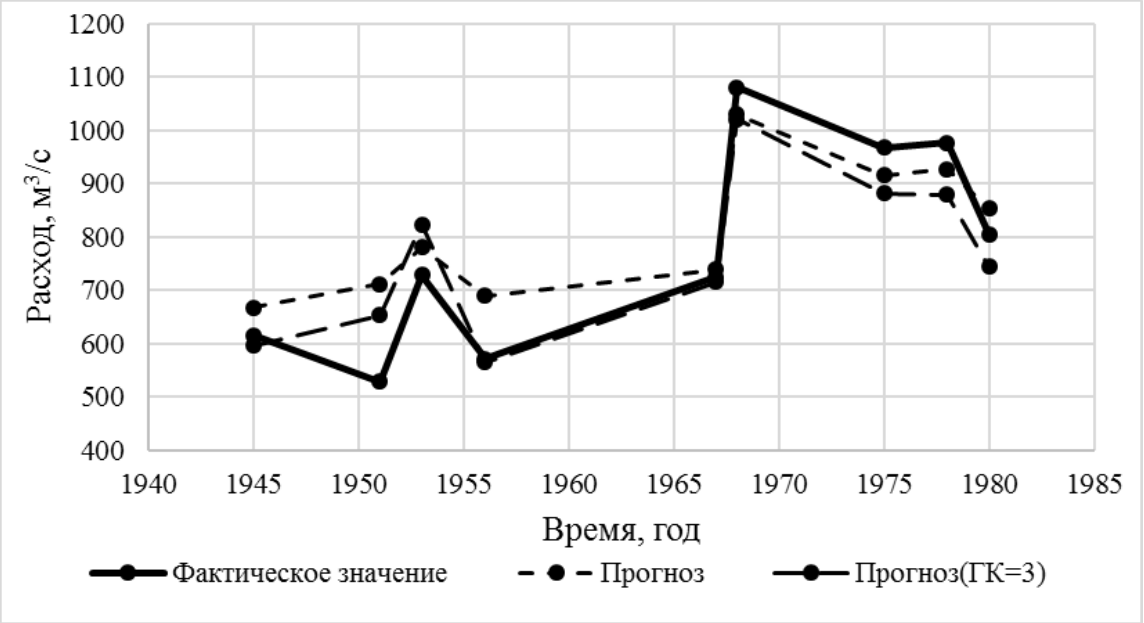


Рисунок 10 – Прогноз значений расхода начала половодья

Прогноз даты пика половодья:

Таблица 14 – Информативность признаков для прогноза даты пика половодья

Все признаки	0,26	0,31	0,22	0,21
ГК=3	0,31	0,37	0,33	
ГК=2	0,53	0,47		

Таблица 15 – Прогноз значений даты пика половодья

Время, год	фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	31 май	5 май	2 май	3 май
1951	16 апр	12 май	8 май	7 май
1953	29 апр	15 май	14 май	19 май
1956	16 май	9 май	24 май	16 май
1967	26 апр	3 май	12 май	8 май
1968	18 май	11 май	9 май	7 май
1975	21 апр	27 апр	30 апр	3 май
1978	24 май	13 май	21 май	24 май
1980	8 май	11 май	8 май	6 май
S / σ		1,07	1,08	1,08

Из таблицы 15 видно, что прогноз не оправдался, видимо, для задачи регрессии даты пика половодья нужно больше входных признаков или больше примеров признакового описания объекта.

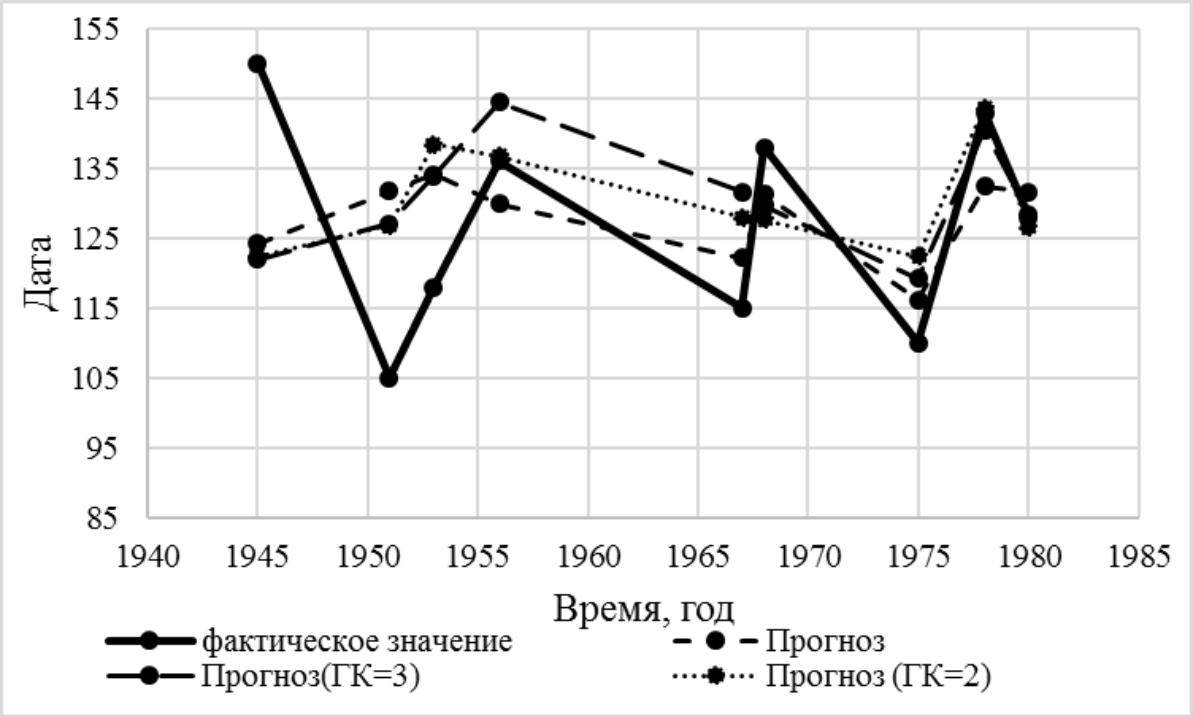


Рисунок 11 – Прогноз даты пика половодья

Прогноз максимального расхода воды за половодье:

Таблица 16 – Информативность признаков для прогноза расхода пика половодья

Все признаки	0,24	0,27	0,32	0,17
ГК=3	0,41	0,35	0,24	
ГК=2	0,52	0,48		

Таблица 17 – Прогноз значения максимального расхода воды за половодье

Время, год	фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	9330	10783,7	10200,8	11174,5
1951	11500	12713,4	12397,1	12697,9
1953	21300	18201,8	15508,7	14218,9
1956	13800	13205,4	12352,5	11571,3
1967	9150	11102,5	11401,4	12338,5

1968	20200	17794,3	17454,4	12122,2
1975	11400	16592,7	17515,3	17841,2
1978	10100	15974,1	15196,2	14218,9
1980	15100	16390,5	14566,8	14993,6
S / σ		0,59	0,67	0,87

Информативность всех признаков достаточно высокая, поэтому сжатие информации привело к снижению точности метода прогноза.

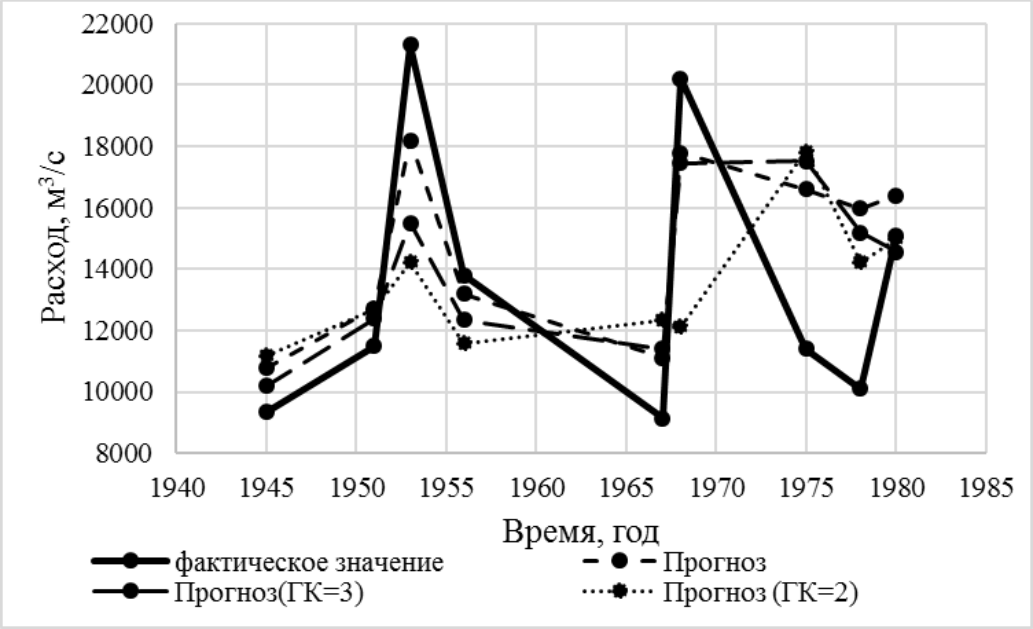


Рисунок 12 – Прогноз значений расхода пика половодья

Прогноз даты окончания половодья:

Таблица 18 – Информативность признаков для прогноза даты окончания половодья

Все признаки	0,20	0,27	0,30	0,23
ГК=3	0,36	0,29	0,35	
ГК=2	0,50	0,50		

Таблица 19 – Прогноз значений даты окончания половодья

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	21 июл	2 июл	3 июл	3 июл
1951	5 июл	9 июл	4 июл	7 июл
1953	30 июн	16 июл	17 июл	10 июл
1956	6 июл	3 июл	10 июл	23 июн

1967	15 июн	12 июн	16 июн	8 июл
1968	28 июл	11 июл	27 июл	10 июл
1975	2 июл	15 июн	15 июн	20 июн
1978	13 июл	26 июн	18 июл	17 июл
1980	1 июл	30 июн	29 июн	3 июл
S / σ		0,65	0,52	0,67

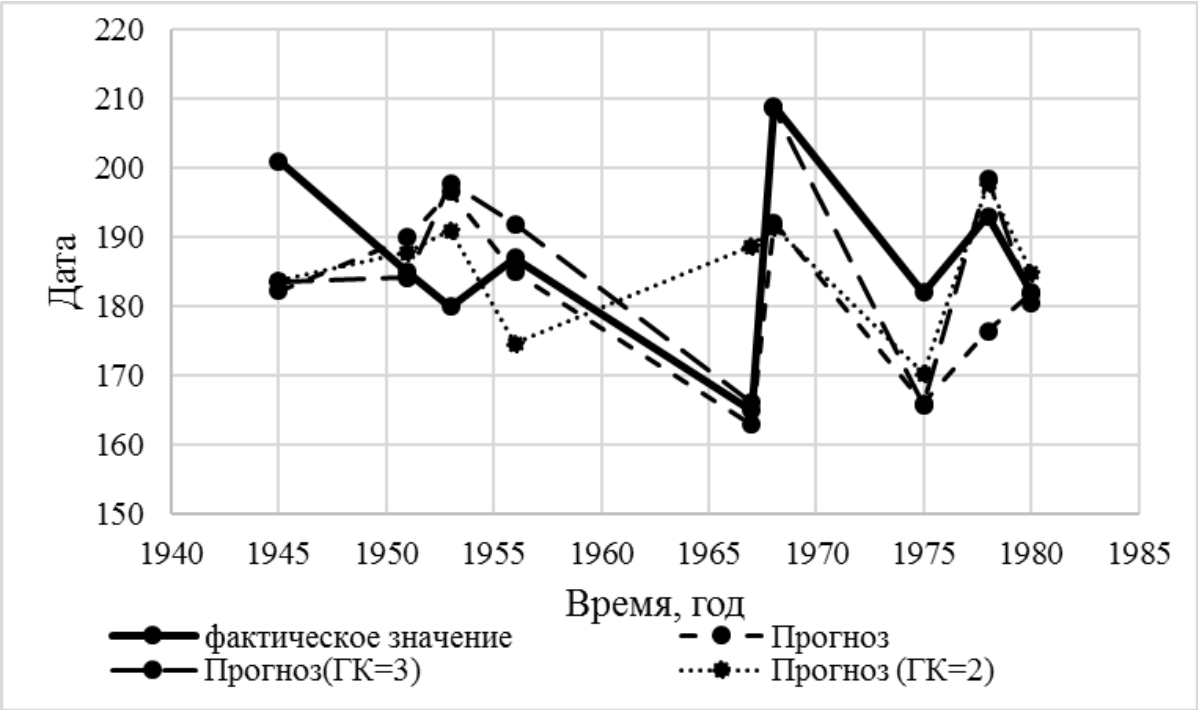


Рисунок 13 – Прогноз значений даты окончания половодья

Прогноз расхода воды окончания половодья:

Таблица 20 – Информативность признаков для прогноза расхода воды окончания половодья

Все признаки	0,24	0,16	0,31	0,29
ГК=3	0,18	0,37	0,44	
ГК=2	0,57	0,43		

Таблица 21 – Прогноз значений расхода воды окончания половодья

Время, год	Фактическое значение	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
1945	1360	2118,3	2540,1	2541,4

1951	2480	2432,7	1994,0	1970,0
1953	2010	2232,6	1653,6	1846,1
1956	1550	2510,9	2696,2	3001,8
1967	1740	2600,6	2439,7	2093,4
1968	2680	1896,9	1512,0	1856,1
1975	2270	2151,8	2089,2	1968,1
1978	3670	2261,0	2248,7	1920,9
1980	1790	2542,1	2321,6	2087,6
S / σ		0,85	0,96	0,99

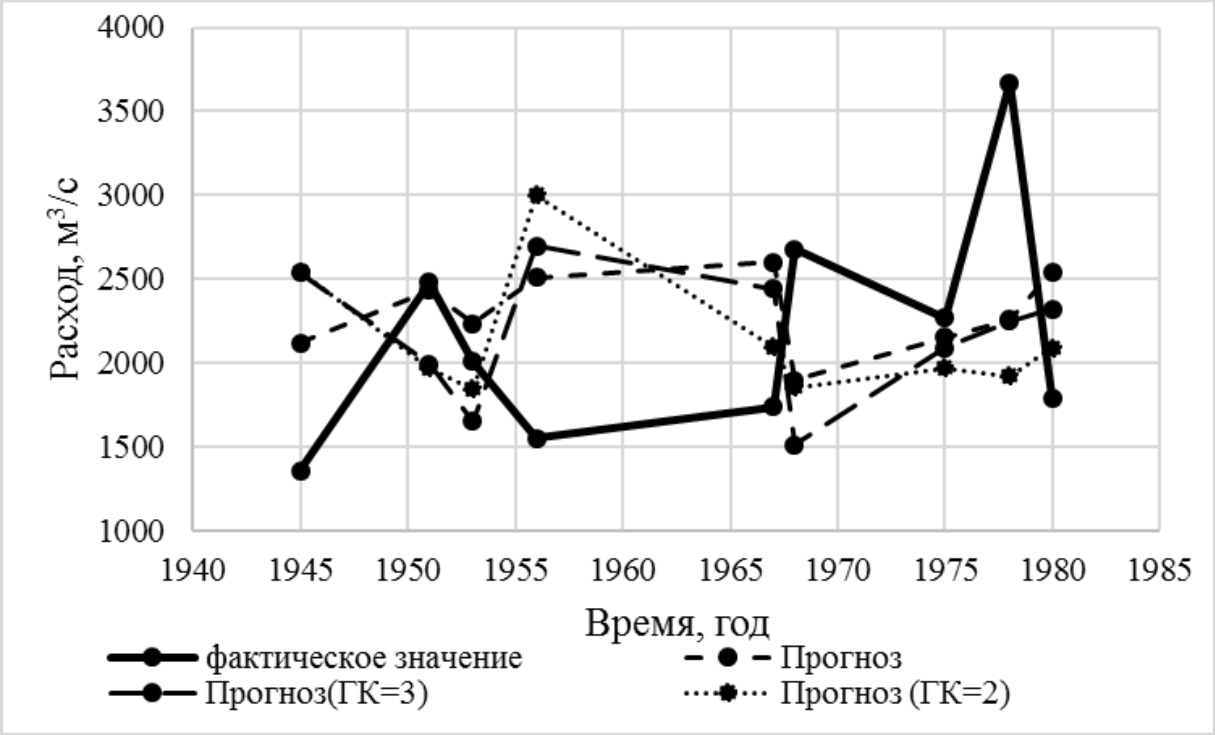


Рисунок 14 – Прогноз значений расхода воды окончания половодья

Критерии качества прогноза превышают допустимое значение. Возможно это связано с недостаточным количеством образов, состоящих из векторов исходных признаков. Или здесь нужно учесть дополнительные признаки, например, спрогнозированные ранее характеристики начала и пика половодья.

Реализация метода в Scikit-Learn следующая:

```
import pandas as pd
import numpy as np
Dat = pd.read_table ('Arg.txt', index_col='God')
target = pd.read_table ('target.txt', index_col='God')
```

```
test = pd.read_table ('test.txt', index_col='God')
from sklearn.ensemble import GradientBoostingClassifier
X = Dat
y = target
clf = GradientBoostingClassifier(n_estimators=100)
clf.fit(X, y)
importances_clf = clf.feature_importances_
pred_clf = clf.predict(test)
#Для регрессии
rg = GradientBoostingRegressor(n_estimators=100)
rg.fit(X, y)
importances_rg = rg.feature_importances_
pred_rg=rg.predict(test)
#Метод главных компонент ГК=3
from sklearn.decomposition import PCA
pca = PCA(n_components=3)
pca.fit(X)
e_v_r=pca.explained_variance_ratio_
comp=pca.components_
Arg_pca=pca.transform(X, y=None)
test_pca=pca.transform(test, y=None)
corr_pca=np.corrcoef(X_pca)
corr_pca=np.corrcoef(test_pca)
```

5.3 Достоинства и недостатки решающих деревьев

Вот некоторые преимущества деревьев решений:

1) Прост для понимания и интерпретации. Деревья можно визуализировать.

2) Требуется мало данных для подготовки. Другие методы часто требуют нормализации данных, необходимо создавать фиктивные переменные и удалять пустые значения.

3) Способность обрабатывать как числовые, так и категориальные данные. Другие методы обычно специализируются на анализе наборов данных, которые имеют только один тип переменных.

4) Способность справляться с проблемами с несколькими выходами.

5) Использует модель белого ящика. Если данная ситуация наблюдаема в модели, объяснение условия легко объясняется булевой логикой. Напротив, в модели черного ящика (например, в искусственной нейронной сети) результаты могут быть труднее интерпретировать.

6) Возможность проверки модели с помощью статистических тестов. Это позволяет учитывать надежность модели.

7) Хорошо работает, даже если его предположения несколько нарушены истинной моделью, из которой были получены данные.

К недостаткам деревьев решений относятся:

1) Ученики дерева решений могут создавать сложные деревья, которые не слишком хорошо обобщают данные. Это называется переобучением. Во избежание этой проблемы необходимы такие механизмы, как, установка минимального количества образцов, необходимых на листовом узле, или установка максимальной глубины дерева.

2) Деревья решений могут быть нестабильными, поскольку небольшие изменения в данных могут привести к созданию совершенно другого дерева. Эта проблема смягчается с помощью деревьев решений внутри ансамбля деревьев.

6 Нейронные сети

Хайкин в своей книге пишет, что нейронная сеть — это громадный распределенный параллельный процессор, состоящий из элементарных единиц обработки информации, накапливающих экспериментальные знания и предоставляющих их для последующей обработки. Искусственная нейронная сеть сильно отличается от биологической. С мозгом её объединяют два принципа:

1) знания поступают в нейронную сеть из окружающей среды и используются в процессе обучения

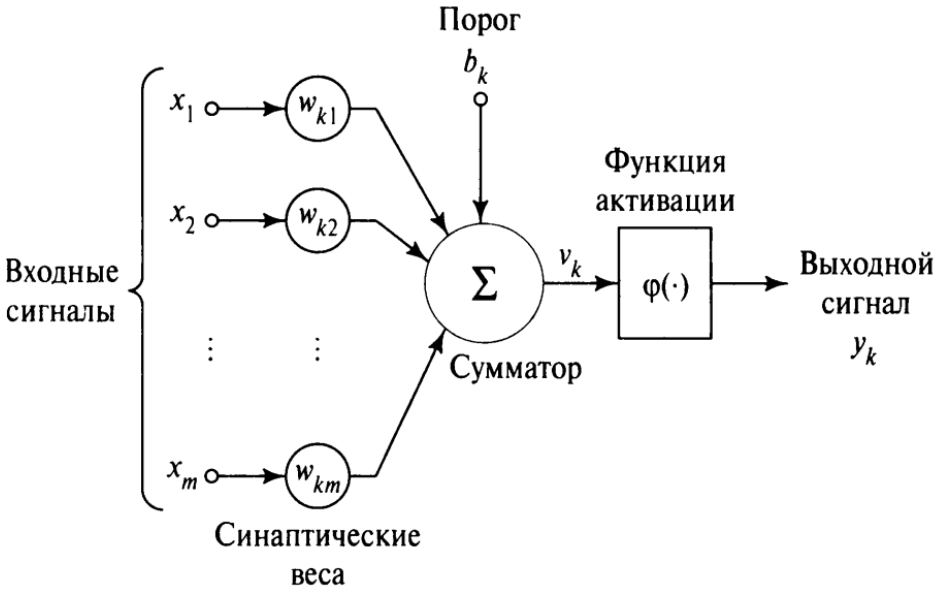
2) для накопления знаний применяются связи между нейронами, называемые синаптическими весами.

Обучить нейронную сеть значит оптимизировать подбор синаптических весов, входящих данных.

Нейронные сети решают задачи обучения как с учителем, так и без него. В задаче кластеризации происходит обучение без учителя, а в задачах классификации и регрессии в большинстве случаев с учителем. Этот метод хорош тем, что нейроны могут быть линейными и нелинейными, большинство зависимостей в природе имеют нелинейный характер, поэтому данный метод должен давать хорошие результаты на обучающей выборке, состоящей из натуральных измерений. Также нейронные сети могут адаптироваться к изменениям внешней среды и обучаться в реальном времени, правда адаптивность не всегда приводит к устойчивости модели.

6.1 Общая модель нейрона

На изображении из книги Хайкена представлена модель нелинейного нейро-



на.

Рисунок 15 – Модель нейрона

Входные сигналы могут быть представлены различными типами данных, в первом слое – это данные обучающей выборки, для остальных слоев – ответы из предыдущего слоя.

Вначале каждой входной точке случайным образом подбирают весовой коэффициент w.

Сумматором называют взвешенную сумму входных точек, с заданным порогом вхождения для увеличения или уменьшения входящих значений.

Функция активации – это функция сжатия информации, в зависимости от её вида можно получить ответ 1 или 0 или число, лежащее в промежутке от нуля до одного или от минус одного до одного.

Функционирование нейрона с номером k можно записать так:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j,$$
$$y_k = \varphi(u_k + b_k) \quad (28)$$

где u_k – линейная комбинация входных воздействий;

$w_{k1}, ..., w_{km}$ – синаптические веса;

$x_1, ..., x_m$ – входные сигналы;

y_k – выходной сигнал;

b_k – порог.

На рисунке 15 v_k – это индуцированное локальное поле нейрона $v_k = u_k + b_k$.

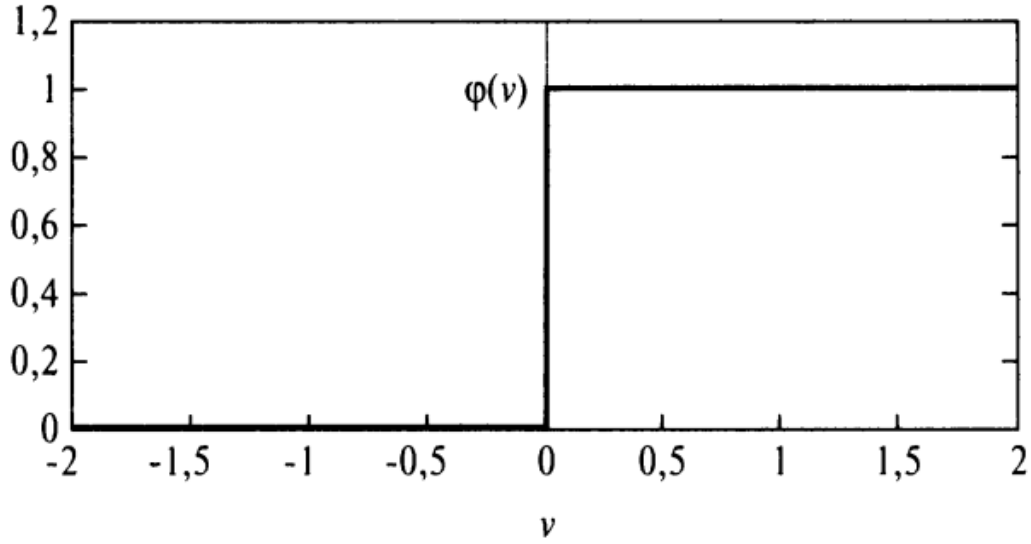
6.2 Типы функций активации

Существует 4 основных типа функции активации:

1) Пороговая

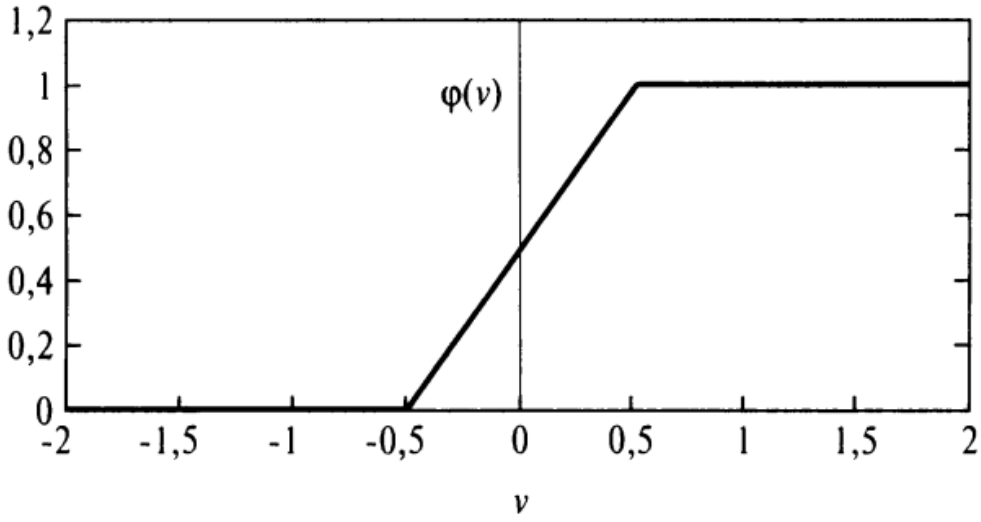
$$\varphi(v) = \begin{cases} 1, \text{ если } v_k \geq 0; \\ 0, \text{ если } v_k < 0; \end{cases} \quad (29)$$

Выходной сигнал, если нейрон активирован, принимает значение один и ноль в другом случае.



2) Кусочно-линейная

$$\varphi(\nu)=\begin{cases}1, & \nu \geq +\frac{1}{2};\\ |\nu|, & +\frac{1}{2} > \nu > -\frac{1}{2};\\ 0, & \nu \leq -\frac{1}{2},\end{cases} \qquad (30)$$



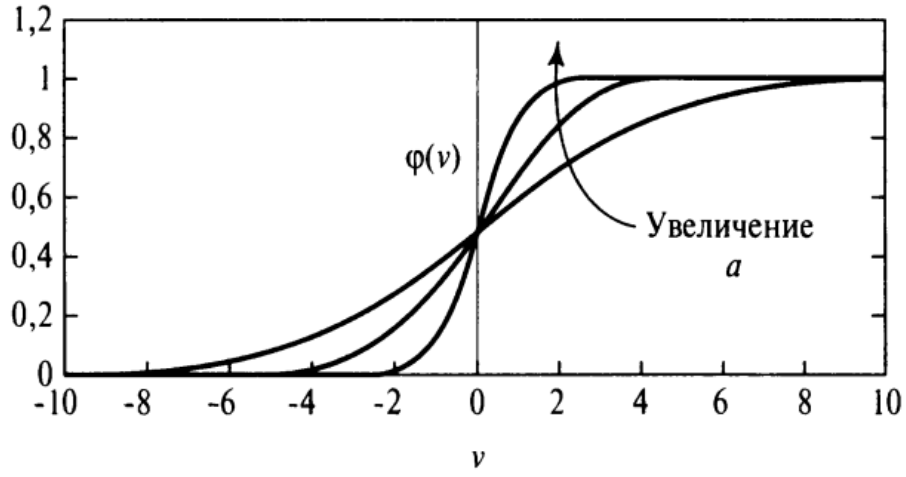
3) Сигмоидальная

Эта функция самая распространённая, так как легко дифференцируема и поддерживает баланс между линейным и нелинейным типом.

Примером такой функции служит логистическая функция вида:

$$\varphi(v) = \frac{1}{1+\exp(-av)}, \tag{31}$$

где a – параметр крутизны или наклона функции.



Ответ такой функции – множество значений от 0 до 1.

4) Функция гиперболического тангенса

$$\varphi(v) = \begin{cases} 1, & \text{если } v > 0; \\ 0, & \text{если } v = 0; \\ -1, & \text{если } v < 0. \end{cases} \tag{32}$$

Как видно из уравнения, она принимает значения от минус одного до одного.

6.3 Стохастическая модель нейрона

Модель нейрона, показанная на рисунке 15, является детерминистской. Это значит, что преобразование входного сигнала в выходной точно определено для всех значений входного сигнала. Однако в некоторых приложениях лучше использовать стохастические нейросетевые модели, в которых функция активации имеет вероятностную интерпретацию. В подобных моделях нейрон может находиться в одном из двух состояний: +1 или —1.

Решение о переключении состояния нейрона принимается с учетом вероятности этого события. Обозначим состояние нейрона символом ж, а ве-

роятность активации нейрона — функцией P(v), где v — индуцированное локальное поле нейрона. Тогда

$$x = \begin{cases} +1, & \textit{с вероятностью } P(v); \\ -1, & \textit{с вероятностью } 1 - P(v). \end{cases} \tag{33}$$

Вероятность $P(v)$ описывается сигмоидальной функцией следующего вида

$$P(v) = \frac{1}{1 + \exp(-v / T)}, \tag{34}$$

где T — это параметр, представляющий эффект синаптического шума.

Заметим, что если параметр T стремится к нулю, то стохастический нейрон, описанный выражением (34), принимает детерминированную форму (без включения шума).

6.4 Нейронные сети прямого распространения

В том случае, если нейроны расположены в несколько слоев, сеть называют многослойной. В случае, когда в сети существует входной слой точек исходных данных, информация от которого передается на выходной слой нейронов, обратной передачи информации не предусматривается, сеть называют сетью прямого распространения или ацикличной сетью. На рисунке 16 изображена структура такой сети, состоящей из четырех нейронов, они же входные, выходные точки. Такая нейронная сеть называется однослойной, единственным слоем принимается слой, содержащий вычислительные элементы (нейроны).

Подсчитывая число слоев сети не учитывается входной слой, так как н не выполняет никаких вычислений.

Если количество слоев больше одного, второй и последующие слои называют скрытыми, а их узлы – скрытыми нейронами. Такие сети относятся к классу многослойных. Скрытые нейроны выполняют посредническую функцию между внешним входом и выходом нейронной сети.

При добавлении дополнительных слоев, появляется возможность выявить сложные статистические связи между входными точками, выделить глобальные свойства, из-за дополнительных синаптических связей и повышения уровня взаимодействия нейронов.

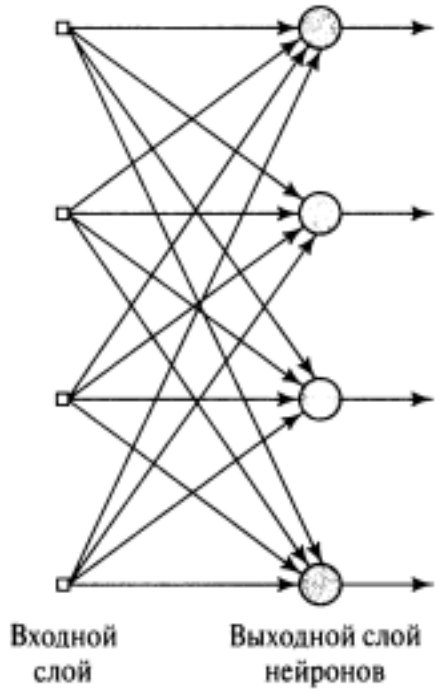


Рисунок 16 – Сеть прямого распространения нейронов

Точки исходных данных, которые образуют входной слой, образуют соответствующие точки шаблона активации, образующий входной вектор данных, для первого слоя, затем шаблон, уже состоящий из выходов первого слоя передается второму слою, по-другому – первому скрытому слою, и так далее, пока не будет получен выход последнего слоя.

Обычно нейроны каждого из слоев сети используют в качестве входных сигналов выходные сигналы нейронов только предыдущего слоя. Набор вы-

ходных сигналов нейронов выходного (последнего) слоя сети определяет общий отклик сети на данный входной образ, сформированный узлами источника входного (первого) слоя. Сеть, показанная на рисунке 17.

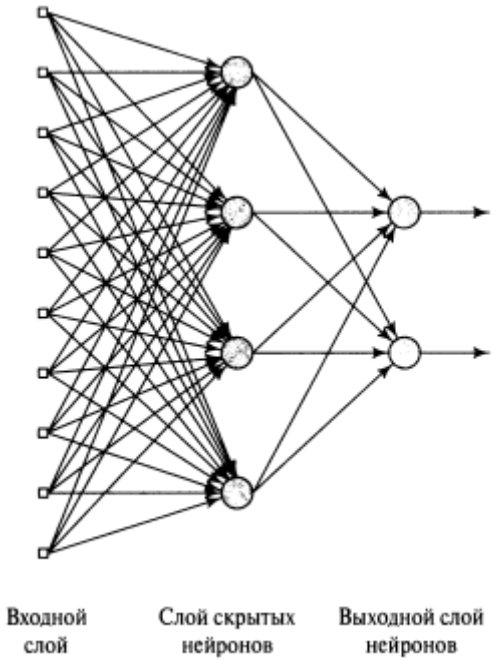


Рисунок 17 – Полносвязная сеть прямого распространения с одним скрытым и одним выходным слоем

Полносвязной сетью можно назвать сеть, в которой все точки одного слоя, соединены со всеми точками смежных слоев, в ином случае сеть называют неполносвязной.

Полносвязные сети прямого распространения ещё называют многослойными персептронами.

Однослойный персептрон, представляет собой простейшую форму нейронной сети, решающую задачу линейной классификации. То есть с его помощью можно классифицировать линейно-разделимые данные. Состоит он всего лишь из одного нейрона, с изменяющимися синаптическими весами и порогом активации, в литературе порог активации ещё называют сдвигом, это название взято из физиологии биологического нейрона.

Алгоритм настройки нейрона однослойной сети был разработан Розенблаттом при попытке создания модели мозга. Он доказал, что если образы (векторы), используемые для обучения персептрона, выбраны из двух линейно-разделимых классов, то алгоритм персептрона сходится и формирует поверхность решений в форме гиперплоскости, разделяющей эти два класса. В исходной версии персептрона, согласно Розенблатту, содержались три типа элементов: сенсорные, ассоциативные и реактивные. Веса связей сенсорных элементов с ассоциативными были фиксированными, а веса связей ассоциативных элементов с реактивными — переменными. Ассоциативные элементы выступали в роли препроцессоров, предназначенных для извлечения модели из данных среды. Что касается переменных весов, то функционирование исходного персептрона Розенблатта в сущности соответствует случаю простого реактивного элемента (т.е. одного нейрона).

Персептрон, содержащий только один нейрон, способен выполнять задачу классификации двух классов. Увеличение числа нейронов персептронной сети приводит к тому, что классификация не ограничивается двумя классами.

Для понимания устройства многослойной персептронной сети, можно ограничиться теорией устройства одного нейрона, который по сути является адаптивным фильтром и работает в линейном режиме.

Многослойные персептроны проходят обучение с учителем, для этого есть несколько алгоритмов, но самым популярным является алгоритм обратного распространения ошибки. Этот метод, как и большинство методов машинного обучения базируется на оптимизации вектора ошибок. Его идея похожа на идею алгоритма минимизации среднеквадратической ошибки, который использовался при градиентном бустинге.

Обучение методом обратного распространения ошибки совершает два обхода по всем слоям сети: прямой и обратный. При прямом проходе образ (вектор признаков) подается на входной слой сети, после чего распространятся по сети от слоя к слою. В результате генерируется набор выходных сигналов, который и является фактической реакцией сети на данный входной образ.

Во время прямого прохода все синаптические веса сети фиксированы. Во время обратного прохода все синаптические веса настраиваются в соответствии с правилом коррекции ошибок, а именно: фактический выход сети вычитается из желаемого (целевого) отклика, в результате чего формируется сигнал ошибки. Этот сигнал впоследствии распространяется по сети в направлении, обратном направлению синаптических связей. Отсюда и название — алгоритм обратного распространения ошибки.

Синаптические веса настраиваются с целью максимального приближения выходного сигнала сети к желаемому в статистическом смысле.

Многослойные персептроны имеют три отличительных признака:

1) Каждый нейрон сети имеет нелинейную функцию активации. Важно подчеркнуть, что данная нелинейная функция является гладкой (т.е. всюду дифференцируемой), в отличие от жесткой пороговой функции, используемой в персептроне Розенблатта. Самой популярной формой функции, удовлетворяющей этому требованию, является сигмоидальная.

2) Сеть содержит один или несколько слоев скрытых нейронов', не являющихся частью входа или выхода сети. Эти нейроны позволяют сети обучаться решению сложных задач, последовательно извлекая наиболее важные признаки из входного образа (вектора).

3) Сеть обладает высокой степенью связности, реализуемой посредством синаптических соединений. Изменение уровня связности сети требует изменения множества синаптических соединений или их весовых коэффициентов.

Комбинация всех этих свойств наряду со способностью к обучению на собственном опыте обеспечивает вычислительную мощность многослойного персептрона.

Однако эти же качества являются причиной неполноты современных знаний о поведении такого рода сетей.

Во-первых, распределенная форма нелинейности и высокая связность сети существенно усложняют теоретический анализ многослойного персептрона.

Во-вторых, наличие скрытых нейронов делает процесс обучения более трудным для визуализации. Именно в процессе обучения необходимо определить, какие признаки входного сигнала следует представлять скрытыми нейронами. Тогда процесс обучения становится еще более сложным, поскольку поиск должен выполняться в очень широкой области возможных функций, а выбор должен производиться среди альтернативных представлений входных образов. Расчет производился в среде разработки MatLab, код программы находится в приложении В.

6.4.1 Расчет характеристик весеннего половодья при помощи многослойного персептрона

Прежде чем преступить к расчету, нужно определиться с количеством скрытых слоев и количеством нейронов в них. В различной литературе, в том числе интернет источниках для поиска сложных закономерностей оговаривают, что трех скрытых слоев вполне достаточно, что касается количества нейронов (h), рекомендуется использовать равное количество нейронов во всех слоях, начинать подбор с h=30, так как с меньшим количеством нейронов результаты почти не меняются, и в зависимости от изменения качества прогнозирования, сокращать шаги подбора количества нейронов в каждом скрытом слое. Еще необходимо задать максимальное количество шагов обучения, в данной работе было задано 20000 шагов (итераций), на каждой итерации корректировались веса нейронов методом обратного распространения ошибки, пока не был достигнут локальный минимум ошибок. В качестве функции активации нейронов была использована сигмоидальная функция.

В задаче классификации для прогнозирования гидрологических характеристик при помощи многослойного персептрона использовались стандартизированные данные, так как ошибки кластеризации на них немного меньше,

по сравнению с кластеризацией нормированных и не масштабированных данных.

Нейронная сеть обучается адаптивно по ходу обучения, подстраиваясь под исходные данные без участия человека, сложно отследить логику принимаемых решений. На выходе получаем модель в виде матриц набора весов для каждого слоя.

В таблицах с результатами серым цветом выделены неправильные ответы значений принадлежности объекта к кластеру. Количество нейронов в каждом слое обозначено буквой h.

Прогноз даты начала половодья:

Таблица 22 – Прогноз значений даты начала половодья, h = 60

Время, год	Фактическое значение	Прогноз
1945	6	6
1951	6	3
1953	7	10
1956	12	12
1967	4	4
1968	15	16
1975	1	1
1978	7	7
1980	6	7

Таблица 23 – Матрица вероятностей принадлежности объекта к одному из кластеров

Ответ сети / Номер кластера	6	3	10	12	4	16	1	7	7
1	0,02	0,00	0,00	0,00	0,00	0,00	0,98	0,00	0,02
2	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,04
3	0,40	0,98	0,00	0,00	0,02	0,00	0,02	0,00	0,03
4	0,00	0,01	0,00	0,01	0,71	0,00	0,00	0,00	0,00
5	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

6	0,62	0,00	0,00	0,00	0,52	0,00	0,00	0,10	0,00
7	0,03	0,01	0,04	0,00	0,00	0,04	0,02	0,15	0,77
8	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
10	0,00	0,01	0,25	0,00	0,00	0,07	0,00	0,11	0,00
11	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00

Продолжение таблицы 23

12	0,00	0,00	0,00	0,94	0,00	0,01	0,00	0,00	0,00
13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
14	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
15	0,00	0,00	0,01	0,02	0,00	0,05	0,00	0,06	0,00
16	0,00	0,00	0,00	0,00	0,00	0,15	0,00	0,01	0,00

Таблица 24 – Прогноз значения даты начала половодья

Время, год	Фактическое значение	Прогноз
1945	01.05.45	22.04.45
1951	02.04.51	07.04.51
1953	08.04.53	24.04.53
1956	29.04.56	24.04.56
1967	10.04.67	07.04.67
1968	14.04.68	12.04.68
1975	09.04.75	07.04.75
1978	10.04.78	05.04.78
1980	24.04.80	21.04.80

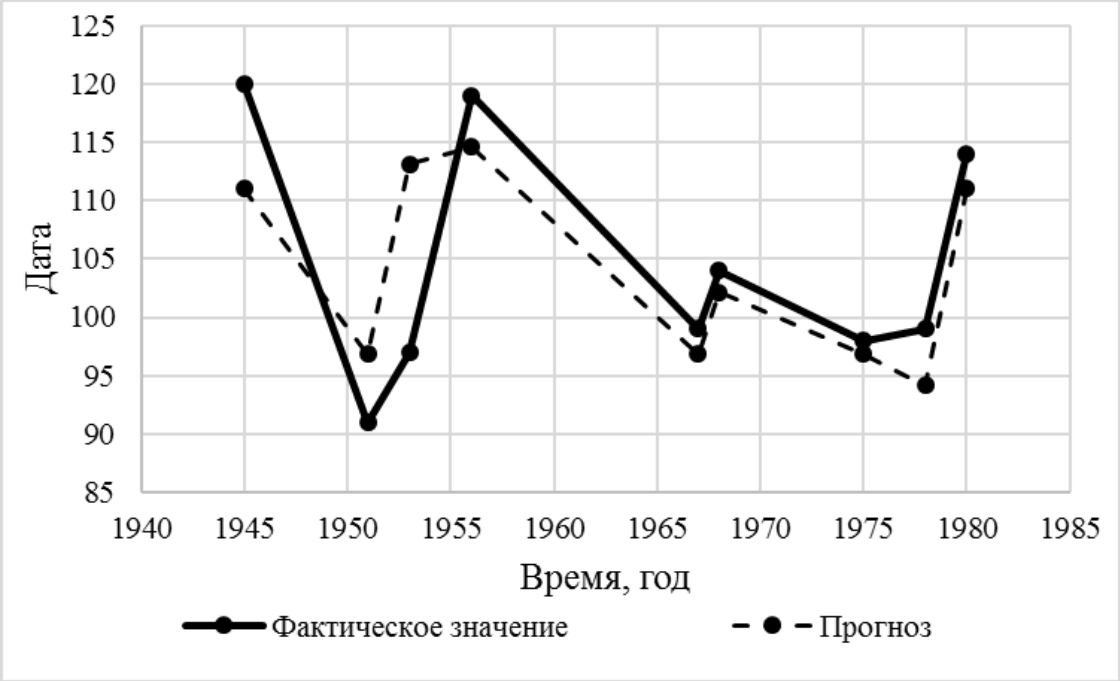


Рисунок 18 – Прогноз значений даты начала половодья

Прогноз расхода воды на начало половодья:

Таблица 25 – Прогноз значения расхода воды на начало половодья h = 60

Фактическое значение	7	7	5	6	7	3	1	2	4
Прогноз	7	7	5	6	7	3	1	2	5

Таблица 26 – Матрица вероятностей принадлежности объекта к одному из кластеров

Ответ сети / Номер кластера	7	7	5	6	7	3	1	2	5
1	0,027	0,012	0,000	0,000	0,010	0,000	0,946	0,001	0,005
2	0,003	0,110	0,001	0,001	0,001	0,099	0,000	0,609	0,195

3	0,000	0,000	0,004	0,008	0,007	0,276	0,000	0,001	0,000
4	0,001	0,038	0,002	0,000	0,597	0,001	0,001	0,037	0,005
5	0,013	0,013	0,999	0,000	0,009	0,072	0,002	0,379	0,395
6	0,001	0,000	0,000	0,996	0,000	0,009	0,000	0,000	0,000
7	0,991	0,836	0,000	0,001	0,839	0,001	0,494	0,000	0,224
8	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000
9	0,000	0,000	0,008	0,001	0,001	0,000	0,001	0,002	0,001

Таблица 27 – Прогноз значений расхода воды на начало половодья

Время, год	Фактическое значение	Прогноз
1945	616	629,67
1951	528	629,67
1953	728	695,09
1956	572	621,14
1967	725	629,67
1968	1080	1072,00
1975	968	1025,14
1978	976	1055,00
1980	805	695,09

Рисунок 19 – Прогноз значения расхода воды на начало половодья

Прогноз даты пика половодья:

Таблица 28 – Прогноз значения даты пика половодья h = 50

Время, год	Фактическое значение	Прогноз
1945	2	2
1951	15	15
1953	5	5
1956	3	3
1967	8	15
1968	6	6
1975	15	15
1978	6	6
1980	5	5

Таблица 29 – Матрица вероятностей принадлежности объекта к одному из кластеров

Ответ сети / Номер кластера	2	15	5	3	15	6	15	6	5
1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,99	0,53	0,01	0,00	0,00	0,00	0,00	0,00	0,32
3	0,00	0,00	0,00	0,41	0,01	0,59	0,00	0,05	0,00
4	0,00	0,00	0,00	0,31	0,00	0,00	0,00	0,00	0,00
5	0,00	0,01	0,98	0,00	0,00	0,10	0,00	0,00	0,76
6	0,00	0,00	0,00	0,00	0,00	0,73	0,00	0,91	0,00
7	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,21	0,00
8	0,01	0,00	0,00	0,01	0,01	0,00	0,33	0,00	0,00
9	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
10	0,00	0,00	0,00	0,00	0,00	0,07	0,00	0,01	0,00

11	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
12	0,00	0,00	0,00	0,00	0,01	0,00	0,02	0,00	0,00
13	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00
14	0,00	0,00	0,00	0,00	0,00	0,00	0,25	0,01	0,00
15	0,01	0,58	0,00	0,00	0,99	0,00	0,43	0,00	0,02
16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Таблица 30 – Прогноз значения даты пика половодья

Фактическое значение	Прогноз
31.05.1945	27.05.1945
16.04.1951	25.04.1951
29.04.1953	06.05.1953
16.05.1956	14.05.1956
26.04.1967	24.04.1967
18.05.1968	15.05.1968
21.04.1975	24.04.1975
24.05.1978	16.05.1978
08.05.1980	06.05.1980

Рисунок 20– Прогноз значения даты пика половодья

Прогноз расхода воды пика половодья:

Таблица 31– Прогноз значения расхода воды пика половодья h = 70

Время, год	Фактическое значение	Прогноз
1945	3	3
1951	3	3
1953	7	5
1956	9	9
1967	3	3
1968	4	4
1975	3	3
1978	5	5
1980	2	2

Таблица 32 – Матрица вероятностей принадлежности объекта к одному из кластеров

Ответ сети / Номер кластера	3	3	5	9	3	4	3	5	2
1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00
2	0,03	0,05	0,01	0,00	0,00	0,00	0,68	0,00	0,97

Продолжение таблицы 32

3	0,98	0,92	0,00	0,00	1,00	0,00	0,84	0,00	0,01
4	0,00	0,00	0,00	0,00	0,00	0,53	0,00	0,00	0,00
5	0,00	0,02	0,71	0,01	0,00	0,28	0,00	0,91	0,12
6	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
7	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
8	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00
9	0,00	0,00	0,00	0,99	0,01	0,04	0,00	0,05	0,00

Таблица 33 – Прогноз значений расхода воды пика половодья

Время, год	Фактическое значение	Прогноз
1945	9330	11148
1951	11500	11148
1953	21300	14522
1956	13800	15880
1967	9150	11148
1968	20200	18629
1975	11400	11148

1978	10100	14522
1980	15100	14545

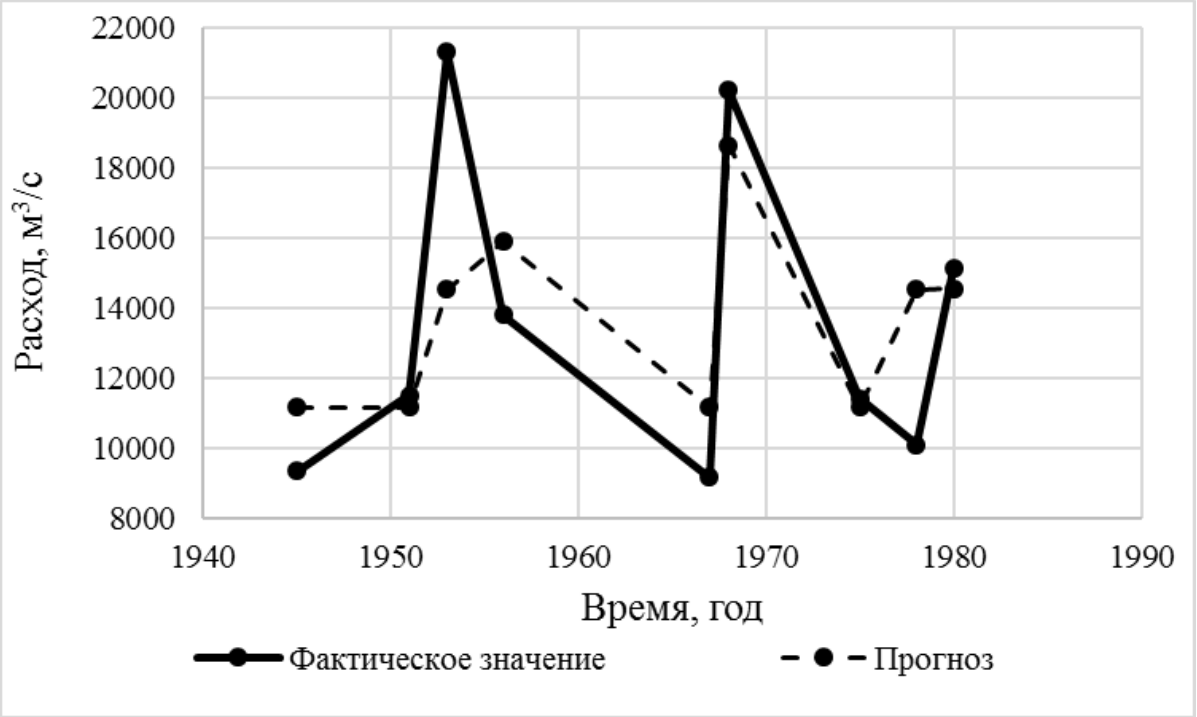


Рисунок 21 – Прогноз значения расхода воды пика половодья

7 Основные результаты

В результате расчетов были получены прогнозные значения дат и величин расхода для начала, пика и окончания весеннего половодья методом градиентного бустинга и прогнозные значения дат и величин расхода для начала и пика половодья методом многослойного персептрона. Решались задачи машинного обучения такие как: кластеризация, классификация и регрессия.

Рассчитана информативность признаков для различных моделей, полученных методом градиентного бустинга. Некоторые признаки несут в себе небольшую информативность, поэтому в расчетах эти признаки можно не учитывать, но так как у нас и так мало признаков, они учитывались, но только с меньшим весом. Уменьшение размерности признаков проводилось при помощи метода главных компонент.

Прогноз даты начала половодья проведен для исходных, нормированных и стандартизированных данных, для выявления влияния преобразования данных на результаты, а также с учетом преобразования данных методом главных компонент.

Оценка качества прогноза методом градиентного бустинга приведена в таблице 33. Здесь характеристика даты начала половодья рассчитывалась при помощи решения задачи классификации, а остальные характеристики – регрессии.

Таблица 33 – Оценки качества метода градиентного бустинга S / σ

Характеристика:	Прогноз	Прогноз (ГК=3)	Прогноз (ГК=2)
Дата начала половодья	0,78	0,71	0,68
Дата начала половодья (по нормированным данным)	0,85	0,73	0,72
Дата начала половодья (по стандартизированным данным)	0,82	0,65	0,94
Расход начала половодья	0,26	0,22	0,53

Дата пика половодья	1,07	1,09	1,09

Продолжение таблицы 33

Расход пика половодья	0,59	0,67	0,87
Дата конца половодья	0,64	0,73	0,64
Расход конца половодья	0,65	0,52	0,67

Известно, что при $n \geq 25$, методика считается эффективной, если $S / \sigma \leq 0{,}80$.

Таким образом, в большинстве случаев методику можно считать эффективной. Учитывая, что данный метод разработан в основном для задачи классификации, погрешности прогнозов можно связать с ошибками этапа разбиения данных на кластеры методом k-средних значений.

Если сравнить ответы кластеров с прогнозируемыми значениями, то видно небольшую разницу значений между правильными ответами и спрогнозированными, в ином случае ошибки классификации могут быть связаны с недостаточным количеством образов объекта, состоящих из векторов признаков, принадлежащих к тому или иному кластеру. Поэтому, можно сказать, что классификация выдала неплохой результат, а прогноз характеристик весеннего половодья – удовлетворительный. Заметим, что сжатие данных влияет на выходные значения, поэтому рекомендуется проверять информативность при разработке модели прогноза, на этапе выбора признакового описание объекта.

Метод многослойного персептрона хорошо справился с задачей классификации, но из-за больших погрешностей кластерного анализа, прогнозы характеристик весеннего половодья получились в основном неудовлетворительными.

Известно, что чем больше образов одного объекта в обучающей выборке, тем точнее прогноз, поэтому те единичные погрешности классификации

методом многослойной сети, скорее всего связаны с недостаточным количеством образов, принадлежащих к одному кластеру.

Приложение

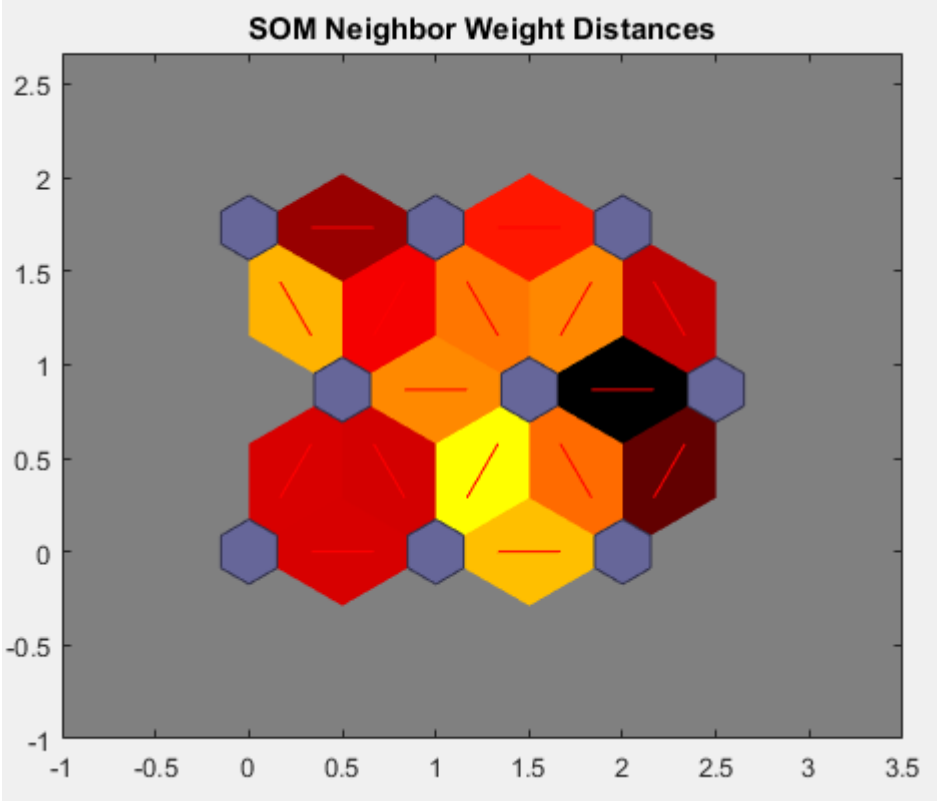
89213054876 Екатерина Владимировна.

Начало половодья

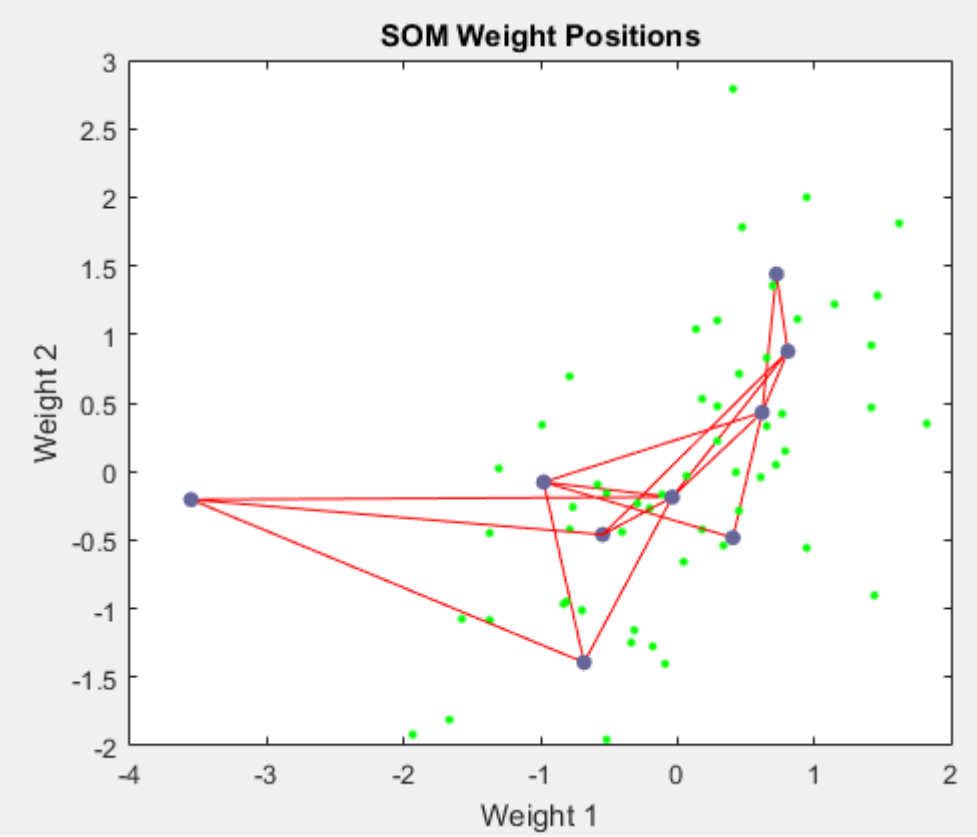
Дата начала половодья

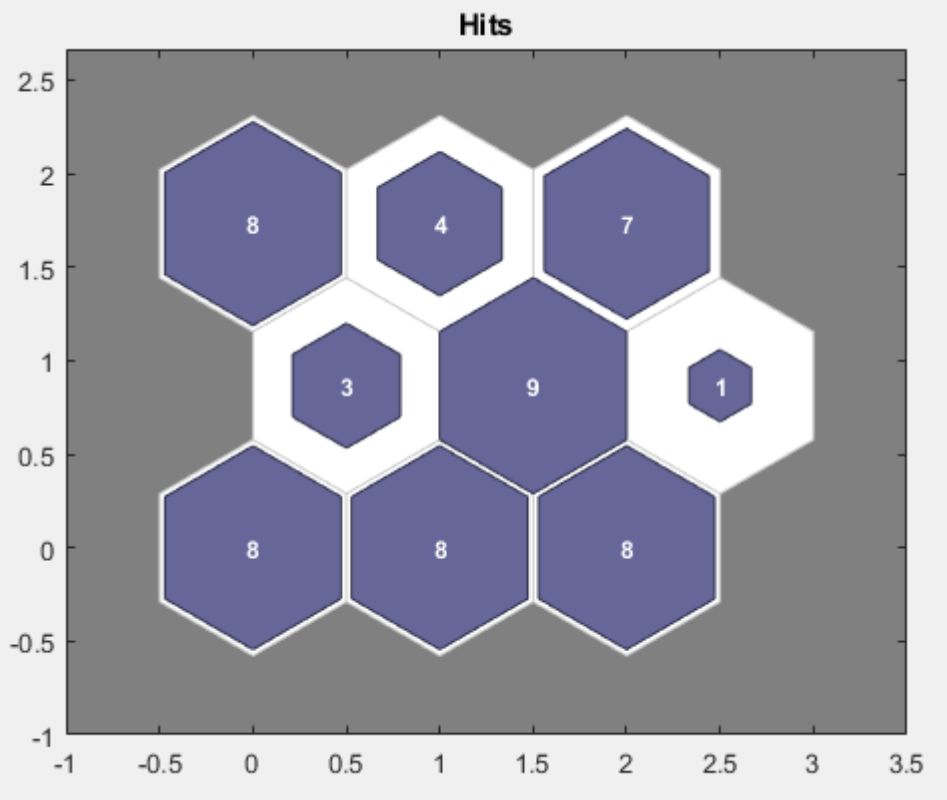
Время, год	Дата	м-д к-средних k=16	ошиб-ка	м-д к-средних k=9 (по нормиро-ванным данным)	ошибка	м-д к-средних k=9 (по стандартизиро-ванным данным)	ошибка
1941	120	12	6,0	9	11,25	1	7,50
1942	104	12	-10,0	6	-8,00	1	-8,50
1943	96	3	0,7	2	-4,33	6	-0,86
1944	110	1	10,6	7	1,80	5	4,86
1945	120	6	16,3	2	19,67	4	9,00
1946	117	15	8,3	9	8,25	1	4,50
1947	104	14	2,0	8	-5,09	7	-9,13
1948	99	7	-3,4	5	-5,14	5	-6,14
1949	99	1	-0,4	7	-9,20	5	-6,14
1950	92	6	-11,7	2	-8,33	6	-4,86
1951	91	6	-12,7	2	-9,33	6	-5,86
1952	110	5	-6,5	7	1,80	7	-3,13
1953	97	7	-5,4	5	-7,14	8	-5,11
1954	102	3	6,7	2	1,67	6	5,14
1955	111	10	3,0	8	1,91	8	8,89
1956	119	12	5,0	6	7,00	2	4,40
1957	116	13	0,3	8	6,91	7	2,88
1958	112	13	-3,7	8	2,91	8	9,89
1959	109	6	5,3	8	-0,09	5	3,86
1960	106	15	-2,7	9	-2,75	4	-5,00
1961	123	5	6,5	7	14,80	7	9,88
1962	99	9	0,0	7	-9,20	8	-3,11
1963	108	15	-0,7	3	-2,50	2	-6,60
1964	115	11	6,0	9	6,25	4	4,00
1965	111	8	0,0	3	0,50	3	0,00
1966	96	16	0,0	9	-12,75	8	-6,11
1967	99	4	-6,0	3	-11,50	6	2,14
1968	104	15	-4,7	9	-4,75	8	1,89
1969	109	12	-5,0	6	-3,00	1	-3,50
1970	101	4	-4,0	2	0,67	9	6,83
1971	110	7	7,6	5	5,86	4	-1,00

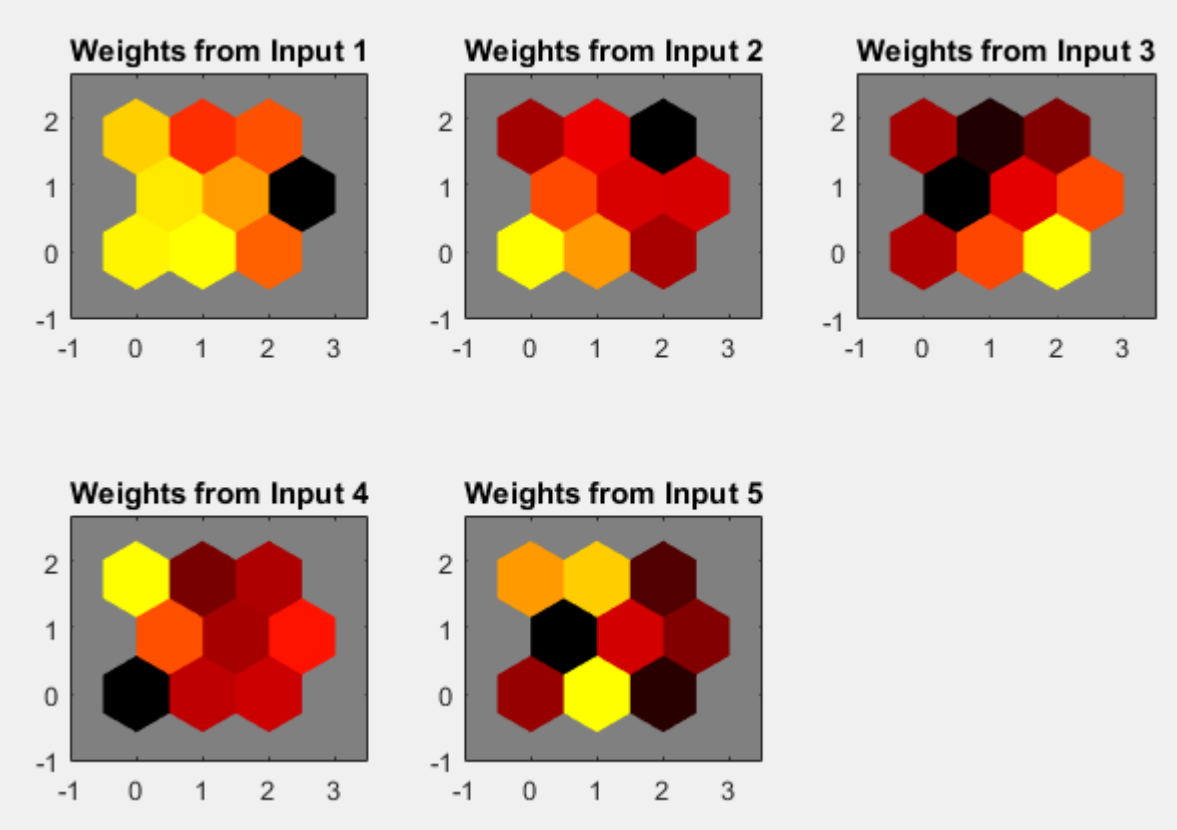
1972	108	11	-1,0	9	-0,75	4	-3,00
1974	117	10	9,0	8	7,91	7	3,88
1975	98	1	-1,4	4	-1,86	6	1,14
1976	109	10	1,0	8	-0,09	7	-4,13
1977	107	7	4,6	5	2,86	4	-4,00
1978	99	7	-3,4	5	-5,14	9	4,83
1979	116	12	2,0	3	5,50	2	1,40
1980	114	6	10,3	5	9,86	4	3,00
1981	119	13	3,3	8	9,91	7	5,88
1982	106	10	-2,0	8	-3,09	5	0,86
1983	90	1	-9,4	4	-9,86	9	-4,17
1984	109	2	6,3	4	9,14	5	3,86
1985	116	12	2,0	6	4,00	2	1,40
1986	103	15	-5,7	5	-1,14	8	0,89
1987	114	15	5,3	3	3,50	2	-0,60
1988	104	11	-5,0	9	-4,75	4	-7,00
1989	104	2	1,3	4	4,14	5	-1,14
1990	88	3	-7,3	1	-4,00	9	-6,17
1991	97	10	-11,0	8	-12,09	8	-5,11
1992	91	2	-11,8	4	-8,86	9	-3,17
1993	107	2	4,3	4	7,14	7	-6,13
1994	100	14	-2,0	8	-9,09	8	-2,11
1995	100	1	0,6	4	0,14	6	3,14
1996	115	4	10,0	3	4,50	4	4,00
1997	96	6	-7,7	1	4,00	9	1,83

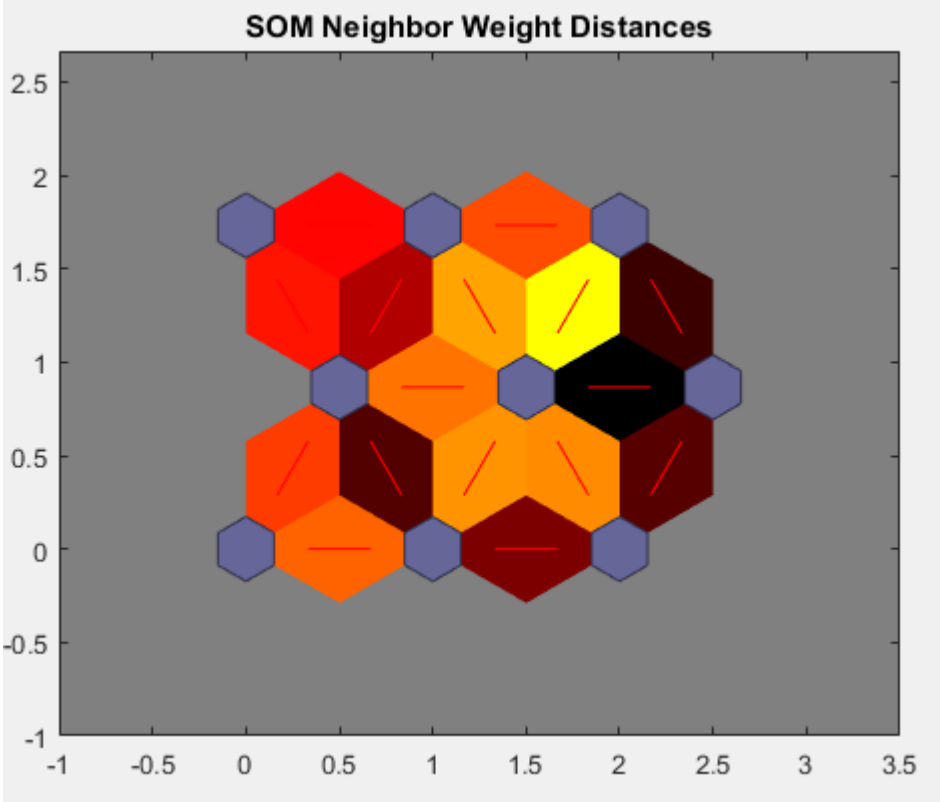


Р-д по стандартизированным данным









Расход на начало половодья

Вре-мя,	Расход начала	м-д k-	ошиб ка	м-д нейро-сетей Кохо-	ошиб ка	м-д нейросетей Кохонена k=9	ошиб ка

год	поло-водья	means k=9		нена k=9 (по норми-рованным данным)		(по стандартизи-рованным дан-ным)	
1941	414	6	-57,1	6	-207,14	3	-6,3
1942	454	6	-115,1	6	-167,14	3	-137,3
1943	696	7	-15,1	7	66,33	9	42,8
1944	464	8	-53,1	8	-143,25	4	53,8
1945	616	7	304,9	7	-13,67	9	-81,3
1946	1020	3	-139,1	3	-52,00	2	5,8
1947	701	5	74,9	5	5,91	1	115,8
1948	672	5	-111,9	5	-23,09	5	6,8
1949	634	8	-74,9	8	26,75	4	-43,4
1950	529	7	91,1	7	-100,67	9	-136,4
1951	528	7	411,1	7	-101,67	9	-283,4
1952	570	8	-62,9	8	-37,25	1	16,6
1953	728	5	-18,9	5	32,91	5	66,6
1954	684	7	-120,9	7	54,33	9	386,6
1955	570	5	-8,9	5	-125,09	4	-43,4
1956	572	6	-103,9	6	-49,14	3	36,6
1957	750	5	-62,5	5	54,91	1	-172,4
1958	927	2	-2,5	2	-128,00	2	-132,4

1959	718	4	17,5	4	-87,67	5	-14,4
1960	565	9	47,5	9	-36,00	3	-21,4
1961	761	8	-61,8	8	153,75	1	-28,4
1962	626	8	8,2	5	-69,09	1	53,6
1963	558	9	25,2	9	-43,00	3	179,6
1964	964	2	43,2	2	-91,00	5	135,6
1965	680	9	-14,8	9	79,00	6	-92,0
1966	780	6	-1,0	6	158,86	2	78,0
1967	725	7	-30,0	7	95,33	9	14,0
1968	1080	3	26,0	3	8,00	2	-115,7
1969	640	6	-132,0	6	18,86	3	-59,7
1970	788	4	48,0	4	-17,67	7	-69,7
1971	1130	2	42,0	2	75,00	2	176,3
1972	744	5	11,0	5	48,91	5	-43,7
1974	1450	2	-42,0	2	395,00	2	-127,7
1975	968	1	66,0	1	-57,14	8	17,3
1976	713	5	12,0	5	17,91	1	242,3
1977	660	5	-207,1	5	-35,09	5	-19,7
1978	976	2	-167,1	2	-79,00	7	0,0
1979	766	6	-49,1	6	144,86	3	-174,9
1980	805	4	158,9	4	-0,67	5	13,1
1981	1020	2	18,9	2	-35,00	2	-44,9
1982	918	2	144,9	2	-137,00	7	-52,9
1983	910	1	100,9	1	-115,1	7	167,1

					4		
1984	1010	1	66,3	1	-15,14	8	367,1
1985	722	6	-13,7	6	100,86	3	-76,9
1986	1100	3	-100,7	3	28,00	2	-197,9
1987	1130	3	-101,7	3	58,00	7	-44,5
1988	1030	2	54,3	3	-42,00	5	-2,5
1989	972	1	95,3	1	-53,14	8	-40,5
1990	1330	1	-147,0	1	304,86	7	87,5
1991	768	5	23,0	5	72,91	5	22,7
1992	886	1	-41,0	1	-139,14	7	-57,3
1993	823	4	150,0	4	17,33	1	-144,3
1994	714	5	15,0	5	18,91	1	-145,3
1995	1100	1	-36,0	1	74,86	8	10,7
1996	935	2	-43,0	4	129,33	9	51,7
1997	765	4	79,0	4	-40,67	7	261,7

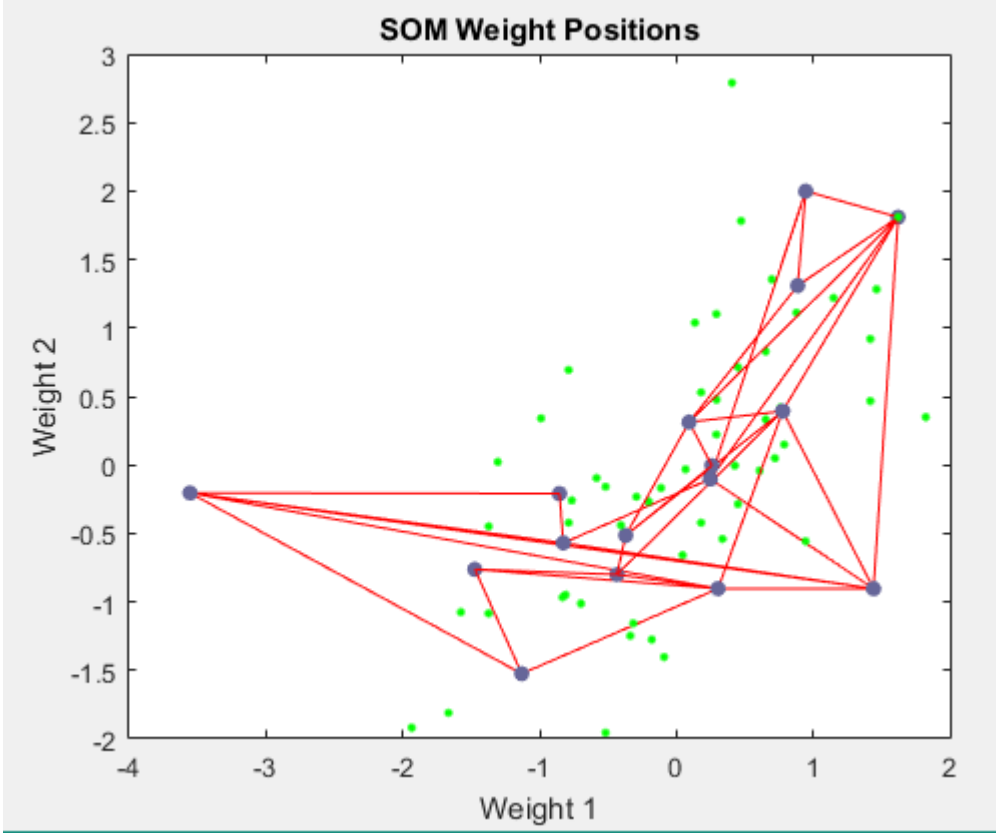
Пик половодья:

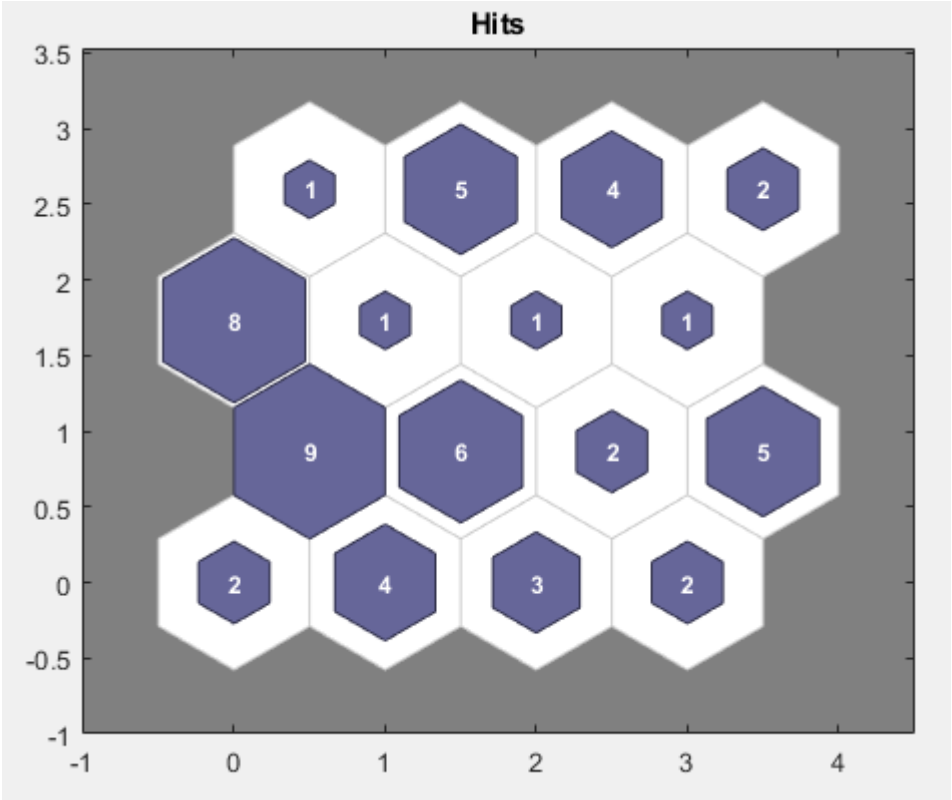
Дата пика по станд.данным

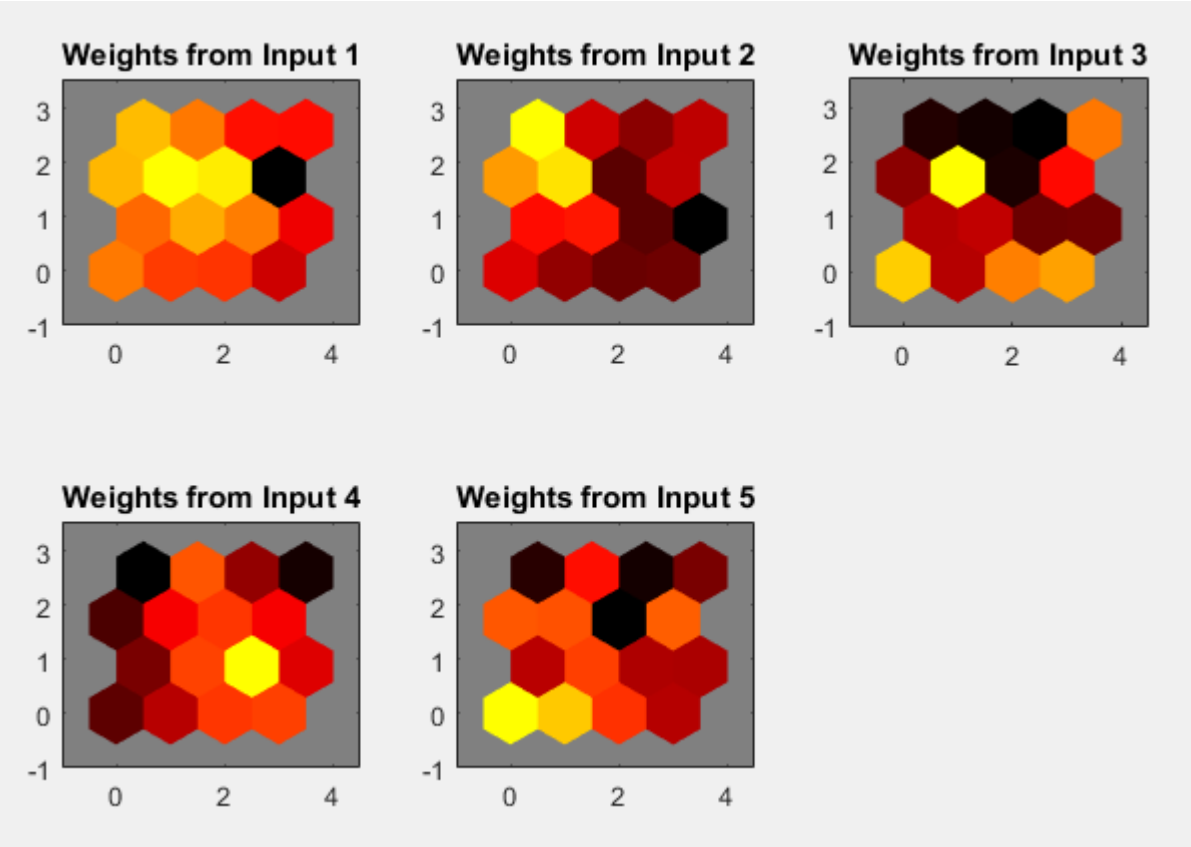
Время, год	Dat_pp	м-д нейросетей кохонена k=16	ошибка	м-д нейросетей кохонена k=9 (по нормированным данным)	ошибка	м-д нейросетей кохонена k=9 (по стандартизированным данным)	ошибка
1941	153	6	16,3	9	17,86	1	3,00
1942	147	6	10,3	6	13,50	1	-3,00
1943	128	7	2,5	2	0,67	8	1,60
1944	138	8	2,0	7	3,60	14	5,80
1945	150	7	24,5	2	22,67	2	4,00
1946	126	3	-6,0	9	-9,14	5	0,22
1947	130	5	0,2	4	-3,55	9	-7,50

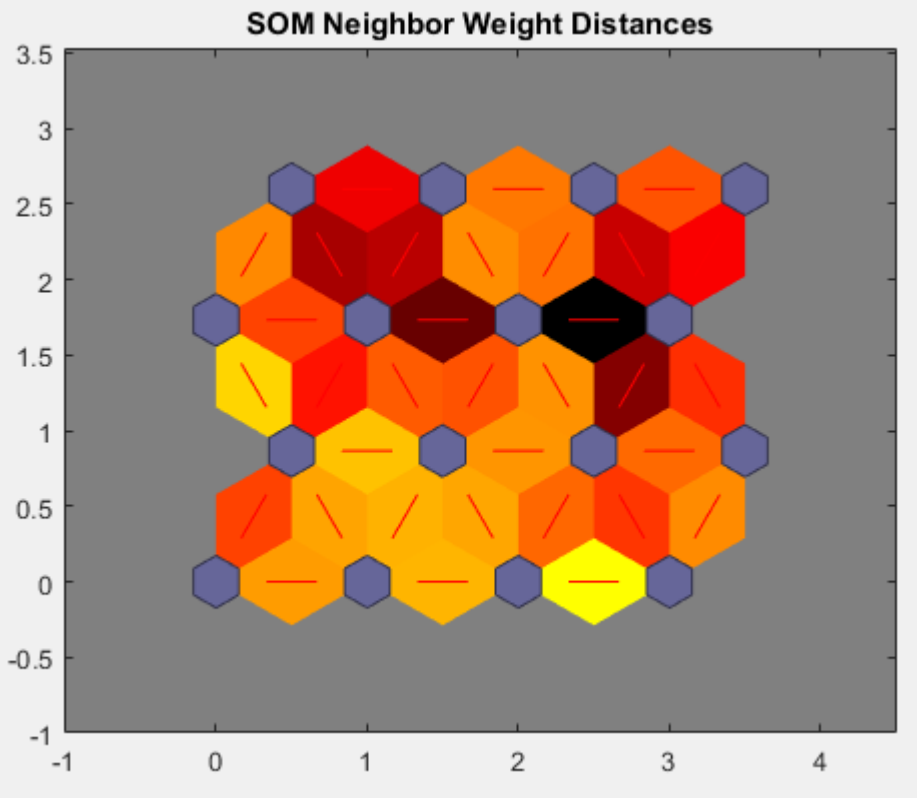
1948	133	5	3,2	5	0,38	6	-2,67
1949	133	8	-3,0	7	-1,40	14	0,80
1950	122	7	-3,5	2	-5,33	8	-4,40
1951	105	7	-20,5	2	-22,33	15	-8,50
1952	158	8	22,0	7	23,60	9	20,50
1953	118	5	-11,8	5	-14,63	5	-7,78
1954	133	7	7,5	2	5,67	8	6,60
1955	134	5	4,2	4	0,45	6	-1,67
1956	136	6	-0,7	6	2,50	3	1,33
1957	133	5	3,2	4	-0,55	9	-4,50
1958	134	2	-6,4	4	0,45	6	-1,67
1959	131	4	1,7	4	-2,55	14	-1,20
1960	119	9	-7,7	9	-16,14	16	-2,00
1961	115	8	-21,0	7	-19,40	13	0,00
1962	128	5	-1,8	7	-6,40	9	-9,50
1963	123	9	-3,7	3	-6,67	4	-2,50
1964	149	2	8,6	9	13,86	2	3,00
1965	138	9	11,3	3	8,33	12	0,00
1966	137	6	0,3	9	1,86	10	0,00
1967	115	7	-10,5	3	-14,67	8	-11,40
1968	138	3	6,0	5	5,38	6	2,33
1969	123	6	-13,7	6	-10,50	16	2,00
1970	126	4	-3,3	2	-1,33	7	1,00
1971	149	2	8,6	5	16,38	2	3,00
1972	126	5	-3,8	9	-9,14	5	0,22
1974	139	2	-1,4	4	5,45	9	1,50
1975	110	1	-10,7	8	-13,71	15	-3,50
1976	142	5	12,2	4	8,45	9	4,50
1977	127	5	-2,8	5	-5,63	5	1,22
1978	143	2	2,6	5	10,38	6	7,33
1979	133	6	-3,7	3	3,33	3	-1,67
1980	128	4	-1,3	5	-4,63	5	2,22
1981	137	2	-3,4	4	3,45	9	-0,50
1982	132	2	-8,4	4	-1,55	6	-3,67
1983	125	1	4,3	8	1,29	14	-7,20
1984	121	1	0,3	8	-2,71	15	7,50
1985	128	6	-8,7	6	-5,50	4	2,50
1986	125	3	-7,0	5	-7,63	5	-0,78
1987	135	3	3,0	3	5,33	3	0,33
1988	136	3	4,0	9	0,86	2	-10,00
1989	125	1	4,3	8	1,29	5	-0,78
1990	112	1	-8,7	1	-6,00	11	0,00
1991	130	5	0,2	4	-3,55	5	4,22
1992	134	1	13,3	8	10,29	14	1,80
1993	133	4	3,7	8	9,29	9	-4,50
1994	127	5	-2,8	4	-6,55	5	1,22

1995	118	1	-2,7	8	-5,71	15	4,50
1996	134	4	4,7	3	4,33	8	7,60
1997	124	4	-5,3	1	6,00	7	-1,00







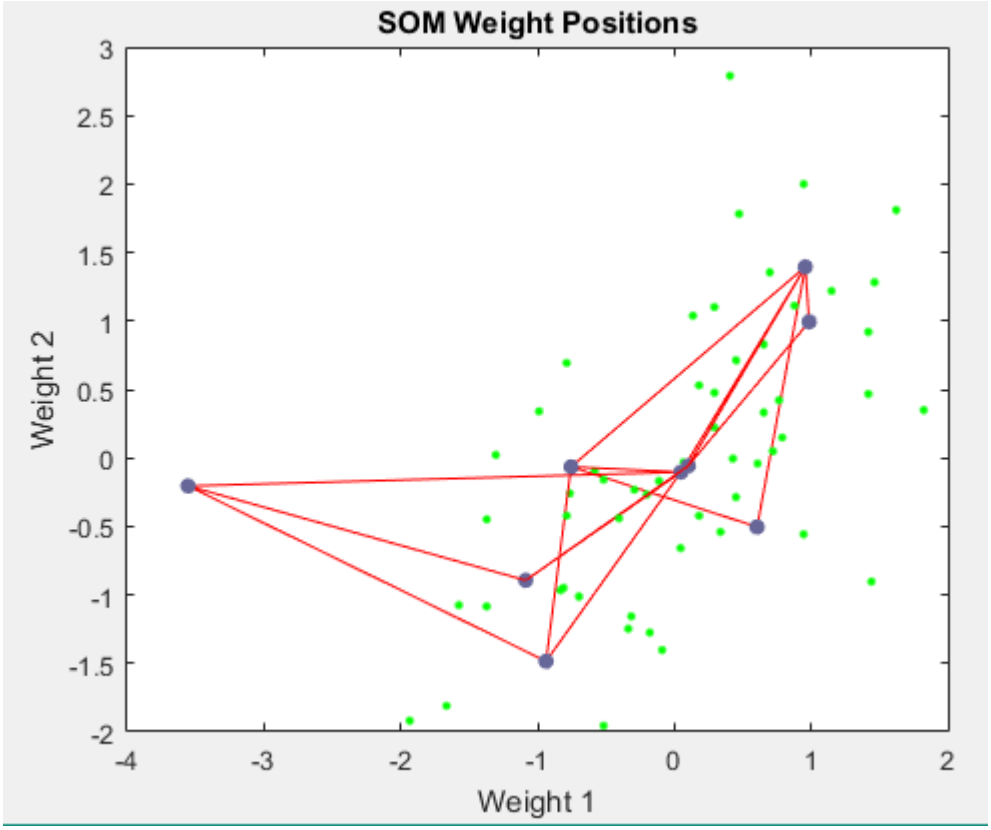


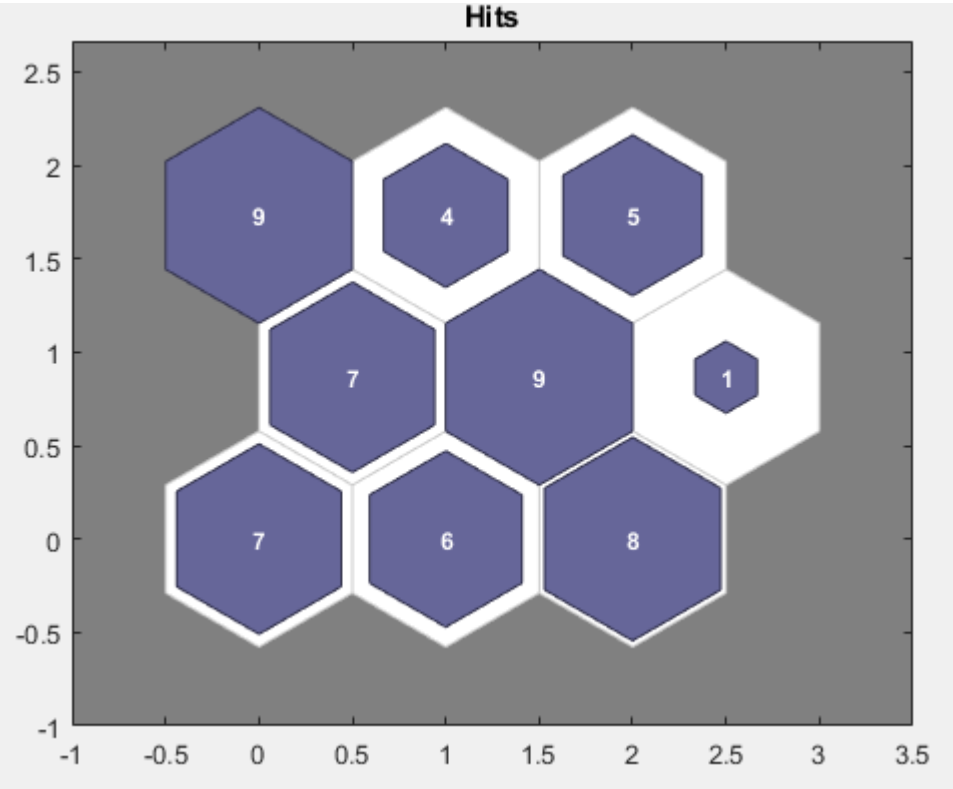
Расход пика

Время, год	Rd_pp	м-д нейросетей кохонена k=16	ошибка	м-д нейросетей кохонена k=9 (по нормированным дан-	ошибка	м-д нейросетей кохонена k=9 (по стандартизированным данным)	ошибка

				ным)			
1941	14500	8	140,0	8	-215,0	8	202,5
1942	12700	12	111,1	16	-850,0	8	-1597,5
1943	12700	12	111,1	14	1195,0	3	1552,5
1944	15200	7	-266,7	9	-2733,3	1	-542,9
1945	9330	16	-340,0	14	-2175,0	3	-1817,5
1946	15990	13	0,0	8	1275,0	8	1692,5
1947	17200	10	200,0	3	-833,3	7	-4044,4
1948	14400	8	40,0	11	-2133,3	5	-122,2
1949	17800	3	-250,0	9	-133,3	1	2057,1
1950	10100	16	430,0	14	-1405,0	3	-1047,5
1951	11500	15	50,0	14	-5,0	3	352,5
1952	17000	10	0,0	1	-1766,7	7	-4244,4
1953	21300	2	50,0	11	4766,7	7	55,6
1954	12600	12	11,1	14	1095,0	3	1452,5
1955	21300	2	50,0	3	3266,7	4	2671,4
1956	13800	14	-33,3	16	250,0	9	-2080,0
1957	24600	1	925,0	6	2280,0	7	3355,6
1958	18200	3	150,0	6	-4120,0	4	-428,6
1959	15700	7	233,3	10	0,0	5	1177,8
1960	9670	16	0,0	8	-5045,0	2	-4875,0
1961	22900	1	-775,0	1	4133,3	7	1655,6
1962	16400	4	-50,0	1	-2366,7	4	-2228,6
1963	17500	6	-33,3	15	2810,0	9	1620,0
1964	14200	8	-160,0	7	-1625,0	5	-322,2
1965	12000	12	-588,9	12	0,0	6	0,0
1966	18700	9	0,0	8	3985,0	4	71,4
1967	9150	16	-520,0	15	-5540,0	3	-1997,5
1968	20200	5	233,3	7	4375,0	4	1571,4
1969	14000	14	166,7	16	450,0	8	-297,5
1970	12800	12	211,1	14	1295,0	1	-2942,9
1971	14600	8	240,0	11	-1933,3	5	77,8
1972	16800	10	-200,0	7	975,0	5	2277,8
1974	24300	1	625,0	6	1980,0	7	3055,6
1975	11400	15	-50,0	2	-3733,3	3	252,5
1976	19400	5	-566,7	3	1366,7	7	-1844,4
1977	15700	7	233,3	11	-833,3	2	1155,0
1978	10100	16	430,0	4	0,0	5	-4422,2
1979	20300	5	333,3	15	5610,0	9	4420,0
1980	15100	7	-366,7	11	-1433,3	2	555,0
1981	22900	1	-775,0	6	580,0	7	1655,6
1982	15200	7	-266,7	3	-2833,3	5	677,8
1983	18100	3	50,0	2	2966,7	1	2357,1
1984	13000	12	411,1	5	0,0	2	-1545,0
1985	13700	14	-133,3	16	150,0	9	-2180,0
1986	18100	3	50,0	11	1566,7	4	-528,6

1987	14100	8	-260,0	15	-590,0	9	-1780,0
1988	12100	12	-488,9	7	-3725,0	5	-2422,2
1989	13000	12	411,1	5	0,0	2	-1545,0
1990	13900	11	0,0	13	-1300,0	1	-1842,9
1991	17600	6	66,7	3	-433,3	5	3077,8
1992	15900	7	433,3	2	766,7	1	157,1
1993	21600	2	350,0	6	-720,0	7	355,6
1994	17500	6	-33,3	3	-533,3	4	-1128,6
1995	20800	2	-450,0	9	2866,7	2	6255,0
1996	12400	12	-188,9	15	-2290,0	3	1252,5
1997	16500	4	50,0	13	1300,0	1	757,1




Hits
9
4
5
7
9
1
7
6
8
2.5
2
1.5
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5